LONG SPAN DNA PAIRED-END TAGS (DNA-PET) FOR UNRAVELING GENOMIC REARRANGEMENTS IN CANCER GENOMES

YAO FEI

(MSc, NUS)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
DEPARTMENT OF EPIDEMIOLOGY AND PUBLIC HEALTH
NATIONAL UNIVERSITY OF SINGAPORE

2011

DECLARATION

I hereby declare that the thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

YAO FEI
27 June 2012

Acknowledgements

Cancer is responsible for one in eight deaths worldwide, however, our understanding of the cancer genomes is still in early days, and frequently, we only identify the tip of the iceberg. In this thesis, I studied few breast and gastric cancer genomes by the long span DNA-PET sequencing technology and revealed some characteristic of cancer genomes. Of course, this thesis would not have been possible without the help of many people. And so, I'd like to thank…

My parents, family, and friends, for supporting me always.

Edison Liu and Ruan Yijun, for being my PhD supervisors, and providing me with a lot of support.

Hillmer Axel, for working together with me on this project and the papers.

Wei Chialin, for giving me all the mentoring and supporting after I joined this group.

Ruan Xiaoan, for giving me all the supporting and caring.

Members of Genome Technology and Biology, especially Audrey Teo and Zhang Zhenshui, the DNA-PET library construction team.

All paper coauthors and people who have contributed to this thesis in one way or another (names are not in any particular order): Herve Thoreau, Melvyn Tan, Yow Jit Sin, Dawn Choi, Low Hwee Meng, Eleanor Wong, Ong Chin Thing (Jo), Neo Say Chuan, Yap Zhei Hwee, Poh Tong Shing, Leong See Ting, Adeline Chew, Jeremiah Decosta, Alexis Khng Jiaying, Lim Kian Chew, Zhang Zhenshui, Audrey Teo, Ruan Yijun, Wei Chia-Lin, Ruan Xiaoan, Edison Liu, Andrea Chavasse, Liu Jun, Patrick Ng, Lee Yen Ling, Jack Tan, James Ye, Lim Yan Wei, Isnarti Bte Abdullah, Guillaume Bourque, Valere Cacheux-Rataboul, Wing-Kin (Ken) Sung, Pramila Ariyaratne, Yanquan Luo, Charlie Lee, Lusy Handoko, Sim Hui Shan, Axel Hillmer, Goh Yu Fen, Christina Nilsson, Zhang Yu Bo, Ngan Chew Yee, Christine Gao, Andrea Ho, Poh Huay Mei, Koichiro Inaki, , Xing Yi Woo, Zhao Hao, Leena Ukil, Jieqi P. Chen, Feng Zhu, Jimmy B.Y. So, Manuel Salto-Tellez, Wan Ting Poh, Kelson F.B. Zawack, Hui Ping J. Lim, Yee Yen Sia, Chee Seng Chan, Patrick B.O. Tan, Atif Shahab, Jonas Bergh, Per Hall, Khay Guan Yeoh, Lance Miller, Chan Yang Sun, Li Guoliang, Melissa Fullwood.

The GIS community, for support and friendly advice.

Publication List

1.Hillmer AM, **Yao F**, Inaki K, Lee WH, Ariyaratne PN, et al. (2011) Comprehensive long-span paired-end-tag mapping reveals characteristic patterns of structural variations in epithelial cancer genomes. Genome Res 21: 665-675. *** I am the co-first author of this paper.**

2. Inaki K, Hillmer AM, Ukil L, **Yao F**, Woo XY, et al. (2011) Transcriptional consequences of genomic structural aberrations in breast cancer. Genome Res 21: 676-687.

3. Ng KP, Hillmer AM, Chuah CT, Juan WC, **Yao F**, et al. (2012) A common BIM deletion polymorphism mediates intrinsic resistance and inferior responses to tyrosine kinase inhibitors in cancer. *Nat Med* **18:** 521-528.

Table of Contents

Summary

All cancers are the result of changes that have occurred in the DNA sequence of the genomes of the cancer cells. We have learned much about these mutations and the abnormal genes that operate in human cancers in the last several decades. Now, we are moving to the new era in which it is possible to get the complete DNA sequence of large number of cancer genomes with the help of the next generation sequencing technologies. In this study, by comparing the characteristics of the large insert libraries (10-20 kb) with short insert (1 kb) libraries with the same sequencing depths and costs, we show that although short insert libraries bear an advantage in identifying small deletions they do not provide a significantly better breakpoint resolution. Large inserts are superior to short inserts in providing higher physical genome coverage and therefore achieve greater sensitivity for the identification of the different types of SVs, including copy number neutral and complex events. Applying the 10 kb DNA-PET technology to 15 cancer genomes, including both primary tumors and cancer cell lines, we show some important genomic characteristics in breast and gastric cancer genomes. With its versatile and powerful nature, the long span DNA-PET sequencing technology has a bright future ahead in studying the structures of numerous normal and cancer genomes.

List of tables

List of figures

List of abbreviations and symbols

array-CGH	array-Comparative Genomic Hybridization
AGH	Array Genomic Hybridization
AML	Acute Myeloid Leukaemia
BAC	Bacterial Artificial Chromosome
BFB Cycle	Breakage-fusion-bridge Cycle
cDNA	Complementary DNA
ChIP-PET	Chromatin Immunoprecipitation with Paired-End Tags
ChIA-PET	Chromatin Interaction Analysis using Paired-End Tag Sequencing
CML	Chronic Myelogenous Leukaemia
CNV	Copy Number Variation
CNA	Copy Number Aberration
cPET	Concordant PET
CTCs	Circulating Tumor Cells
DNA	Deoxyribonucleic Acid
dPET	Discordant PET
DNA-PET	Genomic DNA analysis with Paired-End Tags
EGFR	Epidermal Growth Factor Receptor
FISH	Fluorescence *In-Situ* Hybridization
HD	Homozygous Deletions
IT	Isolated translocation
MAPH	Multiplex Amplifiable Probe Hybridization
M-FISH	Multiplex-FISH
MLPA	Multiplex Ligation-dependent Probe Amplification
NAHR	Nonallelic homologous recombination
NHEG	Nonhomologous end-joining
NGS	Next Generation Sequencing
PARE	Personalized Analysis of Rearranged Ends
PCR	Polymerase Chain Reaction
PEM	Paired End Mapping
PES	Paired End Sequencing
PET	Paired-End Tag
RNA	Ribonucleic Acid
SCLC	Small-Cell Lung Cancer
SKY	Spectral Karyotyping
SNPs	Single Nucleotide Polymorphisms
SNS	Single Nucleus Sequencing
SVs	Structural Variantions
t-AML	therapy-related Acute Myeloid Leukemia
TNBC	Triple-Negative Breast Cancer
TD	Tandem Duplication
TFBS	Transcription Factor Binding Site(s)
UI	Unpaired Inversion

Chapter One: Human cancer genome

Introduction

Cancer is accountable for one in eight deaths worldwide. It ranks as the second leading cause of death in economically developed countries (the first one is heart disease) and the third leading cause of death in developing countries (following heart diseases and diarrhoeal disease) (Garcia M. 2007). It encloses more than 100 different diseases with diverse risk factors and epidemiology which derive from most cell types and organs of the human body. Cancer is characterized by uncontrolled proliferation of cells which can invade beyond normal tissue boundaries and metastasize to distant organs.

(Stephens et al. 2012)In the late nineteenth and early twentieth century, David von Hansemann and Theodor Boveri examined dividing cancer cells under the microscope and observed the presence of bizarre chromosome aberrations. They speculated the numerical alterations on the chromosome level of a cell might contribute to tumourigenesis (Stratton et al. 2009). However, 50 years had been taken to support this speculation by identifying the first specific chromosomal abnormality, a marker chromosome called the Philadephia chromosome in cancer genomes of patients with chronic myelogenous leukaemia (CML) (Rowley. 1973). Subsequently, increased refined analyses of cancer chromosomes showed that cancer genomes are frequently altered in their gross chromosomal structures, and these changes lead to six essential alterations in cell physiology that collectively dictate malignant growth. These alterations include self-sufficiency in growth signals, insensitivity to growth-inhibitory (antigrowth) signals, evasion of programmed cell death (apoptosis),

limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis (Hanahan et al. 2000).

Cancer development is based on two constituent processes, natural selection on the resultant phenotypic diversity and continuous acquisition of heritable genetic variation in individual cells by random mutations. The natural selection may remove cells which have acquired deleterious mutations or it may nourish cells which carry changes that confer the capability to proliferate and survive more effectively than other cells. Like all other cells in the human body, a cancer cell is also a direct descendant, through a lineage of mitotic cell divisions from the fertilized egg, from which the cancer patient developed and consequently carries a copy of its diploid genome. However, there are a set of acquired differences from its progenitor fertilized eggs in cancer cell genomes. These changes are accordingly termed somatic mutations to distinguish from germline mutations which are inherited from parents and transmitted to offspring.

Each somatic mutation in a cancer genome, no matter what kind of structural variations (SVs) it belongs to, may be classified by its consequences for cancer development. 'Driver' mutations have growth advantages on the cells which carry them and have been positively selected during the cancer revolution. They are inherent in the cancer genes. The other mutations are 'passengers' which do not have growth advantages and therefore do not have contributions to cancer development. Passenger mutations in cancer genomes are random somatic mutations without functional consequences. The most important goal of cancer genome research is to identify cancer genes which carry driver mutations and the key challenge therefore is to distinguish between driver and passenger mutations. The general strategy is to uncover a number of structural signatures related with mutations that are under

positive selection. For instance, driver mutations cluster in a subset of genes while passenger mutations are randomly distributed. Currently, at least 384 (2%) of the ~22,000 protein-coding genes in the human genome are reported to have recurrent somatic mutations which contribute to cancer development (Santarius et al. 2010).

Technologies for detecting cancer genomic SVs

Somatic mutations in cancer comprise single nucleotide polymorphisms (SNPs) and SVs which can be further divided into microscopic variations (larger than 3 Mb) and submicroscopic variations. SVs include deletions, insertions, inversions, duplications, translocations and copy number variations (CNVs) (Feuk et al. 2006; Sharp et al. 2006). All these SVs can be separated into two groups: balanced and unbalanced rearrangements. Based on the size and the pattern of SVs, different kinds of techniques can be chosen to interrogate the rearrangements in cancer genomes.

1.1 Cytogenetic methods

Since its first application in cancer research, cytogenetics has taken us from a state of nearly no knowledge of the chromosome changes in human cancer to a point at which an incredible body of information is available. The first era of cytogenetics began in 1956 when the techniques were developed to accurately determine the number of chromosomes in a cell. However, early cytogenetics did not allow to distinguish most individual chromosome pairs from others of similar sizes and to define their general morphology. The chromosome basis of many diseases such as Down syndrome, Turner syndrome, Klinefelter syndrome had been established by these techniques; however, segmental gains or losses smaller than 25-50Mb in size were generally not detectable by these methods (Friedman. 2009). The second era of cytogentics started at the beginning of 1970s as the development of techniques for banding of human

chromosomes emerged. These techniques provide the ability to distinguish every chromosome pair unequivocally and identify deletions and duplications as small as 5-10 Mb. The second era of cytogenetics led to the recognition of a large number of micro deletion/ micro duplication syndromes, including the Aniridia-Wilms tumor syndrome and those related with terminal deletion of chromosome 11q.

The introduction of fluorescence in situ hybridization (FISH) and subsequently stretched-fiber FISH marked the beginning of another new era of cytogenetics. By using fiber FISH, the resolution can be improved from the whole chromosomes in metaphase at ~5 Mb, or interphase nuclei at 50 kb to 2 Mb to the level of chromatin strands (5-500 kb) (Speicher et al. 2005). Although FISH is a robust test that can be used in a variety of diagnostic applications (Volpi et al. 2008), FISH only provides information about one or a few specific genomic regions which have been chosen for testing. Genome wide FISH based methods which include chromosome painting, Multiplex-FISH (M-FISH), spectral karyotyping technique (SKY) and cytogenetic comparative genomic hybridization are useful to clarify the nature of some chromosomal markers and rearrangements, but the resolution is still relatively low (Xu et al. 2003).

1.2 Array based method.

In the past decade, a revolutionary technology called array-comparative genomic hybridization (array-CGH) or more generally, array genomic hybridization (AGH) permit the identification of submicroscopic losses or gains anywhere in the genome. With thousands of locus specific probes immobilized onto microarrays, the patient and healthy reference genomic DNA can be compared by hybridization (Sharp et al. 2006). Genomic clones (for instance, Bacterial Artificial Chromosome, BAC),

cDNAs, PCR products or oligonucleotides had been used as microarray probes. Among of them, BAC and oligonucleotide arrays are the most widely used probes. The advantage of AGH compared to conventional cytogenetic methods is the resolution as it can detect chromosomal gains or losses as small as 50-100kb anywhere in the genome. This is 100 times smaller than the resolution of the traditional cytogenetic method.

In the past few years, genome-wide genotyping arrays have been developed and applied in CNV detection. For instance, Affymetrix SNP Array 6.0 has 1.8 million genetic markers, representing more than 906,600 SNPs and 946,000 probes for CNVs. Meanwhile, Illumina Human1M BeadChip features more than 1.07 million SNP markers for CNV analysis covering 14,000 total CNV regions (Wu X M. 2009). Although array-based methods provide a genome-wide screening of variations, they are unable to detect copy number neutral variants, such as inversion or balanced translocation. They also cannot precisely describe the breakpoints and other fine details of the genomic rearrangements.

1.3 PCR based methods

Multiplex amplifiable probe hybridization (MAPH) and multiplex ligation-dependent probe amplification (MLPA) are two methods developed based on quantitative PCR. In 2000, MAPH was first described (Armour et al. 2000) while two years later, MLPA was developed (Schouten et al. 2002). Both technologies rely on comparative quantification of specifically bound probes that are amplified by PCR with universal primers. In the MAPH technique, low amount genomic DNA (1μg) without any pre-amplification is fixed onto a membrane and hybridized with a set of probes corresponding to the target sequences to be detected. All probes are flanked by the

same sequence. After removing unbound probes by stringent wash, the remaining specifically bound probes are then stripped from the membrane and amplified by the universal primer pair. Since each amplified probe has a different length, the PCR products can be separated by gel electrophoresis and the amount of amplified product is directly proportional to the copy number in the genomic DNA (Sellner et al. 2004). QuadMAPH is a newly developed method which allows the user a four-fold increase in the number of loci tested simultaneously without the expense of genome-wide approaches (Tyson et al. 2009). Duet to its more limited DNA consumption (50ng *per* assay) and single-tube assay set up, MLPA has become widely used compared to MAPH. In the MLPA technique, each target has two probes which hybridize adjacently to each other and all probe pairs are flanked by universal primers. After hybridization, the two parts of the probe pair are joined together by a ligation reaction, and the number of ligated products is proportional to the target copy number. Currently, MAPH and MLPA fill an important gap in the assay methodology between genome-wide CGH platforms and single locus low-throughput methods. The most important advantages of these two methods are relative simplicity, speed, cost-effectiveness and high accuracy in detecting small genomic changes; however, the limited complexity due to gel-based detection is a major drawback (Kozlowski et al. 2008).

1.4 Sequence based methods

The complete sequence of the human genome (International Human Genome Sequencing Consortium, 2004) provides another approach to detect SVs. In 2005, Tuzun et al. sequenced a high density fosmid library by conventional sequencing technology (ABI3100) and mapped 589,275 paired end sequences against the human reference genome assembly (Tuzun et al. 2005). Around three hundred sites of SVs

had been identified including 139 insertions, 102 deletions and 56 inversion breakpoints. This method can simultaneously sequence and precisely characterize any structural variation based on a clone library. It allows the detection of balanced rearrangements like inversion. However, due to the insert size (~40 kb) of the fosmid vector, the smallest size of detectable SVs is larger than 8 kb in length. And the high cost of the conventional sequence technology also prohibits this method to be widely used for genome wide SVs detection.

We just experience the next technological revolution in biological science, the advent of relatively cost effective, massively parallel DNA sequencing technologies. Over the last two years, it has become possible to resequence the entire human genomes at single nucleotide resolution. It is expected that in the next 3-4 years, a single human genome can be sequenced within days for around 1,000 dollars (Aparicio et al. 2010).

The most limited step in conventional Sanger-based sequencing of DNA comes from separating randomly terminated DNA polymers by gel electrophoresis. Next generation sequencing (NGS) devices bypassed this limitation by physically arraying DNA molecules on solid surfaces and determining the DNA sequence *in situ*. The first generation approaches, for example Roche 454 sequencing (Margulies et al. 2005), currently achieve 200-400 bp reads over hundreds of thousands of templates. The current generation of machines, such as Illumina HiSeq devices (Bentley et al. 2008), ABI SOLiD machines (McKernan et al. 2009) and Complete Genomics Amplified DNA nanoarray platform (Drmanac et al. 2010) are capable of sequencing tens of millions of individual templates in parallel with 100 bp read length. Paired-end sequencing (PES) (Holt et al. 2008), paired- end mapping (PEM) (Korbel et al. 2007) or Paired-end tag (PET) (Chen et al. 2008) are the methods developed based on the

NGS. Albeit different nomenclature, the underlying principle is the same: the extraction of short tag sequence information from the two ends of long DNA fragments, the pairing of the two tags, and the mapping of the paired tag sequences to reference genomes for demarcating the boundaries of the target DNA fragments in the genome landscape (Fullwood et al. 2009). PET can be used on RNA (RAN-PET) for transcripton analysis (Morin et al. 2008), on DNA (DNA-PET) for the identification of genome SV and can aid genome sequence assembly (Korbel et al. 2007). Further, PET can be applied on manipulated DNA fragments such as ChIP-enriched DNA (ChIP-PET) for mapping of transcription factor binding sites (TFBSs) (Wei et al. 2006) and on proximity-ligated DNA for chromatin interaction analyses (ChIA-PET) (Fullwood et al. 2009).

The first genomic sequence of a cancer was described by Ley et al (Ley et al. 2008) and Illumina GA technology was used to attain a nearly 33-fold coverage of an acute myeloid leukaemia (AML) patient genome. At the same time, a 14-fold genome coverage of a normal skin sample from the same patient was obtained as a germ line control. The genome of this patient is cytogenetically normal and diploid, which represent a simpler case than the genomes of cytogenetically complex and much more common carcinomas. However, the analysis provides an informative snapshot of what we can expect from cancer genome resequencing. A total of ten non-synonymous somatic mutations were identified in the patient's genome. Two are well known AML-associated mutations and both of them are common (25-30%) in AML tumors. The other eight are new described and of unknown function or relevance to the tumorigenic process. This study establishes whole-genome sequencing as an unbiased method for discovering cancer-initiating mutations in previously unidentified genes that may respond to target therapies.

The same strategy had been applied to another AML patient and a total of 64 somatic mutations had been identified (Mardis et al. 2009). Four mutations occurred in at least one additional AML tumor sample of 188 additional AML samples that were tested. Mutations in *NRAS* and *NPM1* had been identified before but the other two mutations were new. One mutation in the gene *IDHI* was present in 15 of 187 additional AML genomes and was strongly associated with normal cytogenetic status. The other mutation located in a non-genic evolutionarily conserved region, is pointing to the importance of the development of a strategy for informative analyses of non-protein encoding alterations with potential regulatory functions. Understanding the potential regulatory effects of these alterations will be of key importance in understanding the molecular mechanism of cancer.

The advances in NGS also can be used to characterize all somatic mutations that occur during the development and progression of individual cancers. The genome of an estrogen-receptor-α-positive metastatic lobular breast primary tumor and a metastasis collected from the same patient 9 years later had been sequenced and compared (Shah et al. 2009). The results provided insight into the evolution of the cancer genome associated with disease progression. Of all the 32 somatic mutations detected in the metastasis, only 11 could be detected in the primary tumor. Five of the 11 mutations were prevalent in the DNA of the primary tumor and the others were present at low frequencies (1-13%). The authors noted that the prevalence of new mutations in metastases could reflect those associated naturally with tumor progression, and those induced by treatments such as radiation therapy. Another significant feature of this study is the integration of genome and transcriptome analysis. The authors examined how the transfer of information from the nuclear genome to proteins was modified by alternative splicing, biased allelic expression and

RNA editing. Several hundred putative RNA-editing events were observed that would potentially result in non-synonymous protein changes not coded directly by the gene. Two genes, *COG3* and *SRP9*, showed confirmed high frequency non-synonymous transcript editing, resulting in variant protein sequences. More interesting, the ADAR enzyme, one of the principal RNA-editing enzymes, was the only editing enzyme expressed at a high level. These observations highlight the importance of integrating RNA-seq data with tumor genomes, and that the quantitative and digital aspects of NGS can together be applied to understand gene activation/inactivation.

In the same year, Pleasance and colleagues sequenced two human cancer cell lines, NCI-H209 which is an immortal cell line derived from a patient with small-cell lung cancer (SCLC) and COLO-829 derived from a metastasis of a malignant melanoma patient (Pleasance et al. 2010a; Pleasance et al. 2010b). In the first case, total 22,910 somatic substitutions were identified, including 134 in coding exons. However, most of the mutations in coding and promoter regions of the NCI-H209 genome are passenger events, without selective advantages to the cells. At least two separate DNA repair pathways have been uncovered for protection of the NCI-H209 genome, and the two pathways operated with differing efficacy across six classes of mutation, which implies that the lesions have distinct physicochemical effects on DNA structure, and could be variable recognized and excised by the genome surveillance machinery. A 39 kb tandem duplication (TD) was found in *CHD7*, predicted to lead an inframe duplication of exon 3-8 and another two SCLC cell lines carried a *PVT1-CHD7* fusion gene within the MYC amplification. As a chromatin remodeler, *CHD7* can promote enhancer-mediated transcription through association with histone H3K4 methylation. Histone modifiers have been implicated as cancer genes and the recurrent of *CHD7* in SCLC will extend this theme.

In the second case, a malignant melanoma and a lymphoblstoid cell line from the same person had been sequenced and provided the first comprehensive catalogue of somatic mutations from an individual cancer. Most somatic base substitutions in COLO-829 were C>T/G>A transitions and this indicated that COLO-829 were attributable to ultraviolet-induced DNA damage. The catalog showed traces of the multiple levels of selective application of DNA repair: targeting transcribed rather than untranscribed, exons rather than introns, transcribed DNA strands rather than non-transcribed strands, and 5' rather than 3' ends of genes.

Tumor-specific chromosome rearrangements have the potential to serve as highly sensitive biomarkers for tumor detection, such as those involving the *BCR-ABL* oncogene (Hughes et al. 2006). Rearrangement-associated biomarkers also offer a reliable measure that would be useful for monitoring tumor response to specific therapies, detecting residual disease after surgery and long-term clinical management. However, recurrent structural mutations do not generally occur in most solid tumors. A method, called personalized analysis of rearranged ends (PARE) had been developed to identify translocations in solid tumors (Leary et al. 2010). Four colorectal and two breast cancers had been sequenced by massively parallel sequencing and revealed an average of nine rearranged sequences (range, 4 to 15) per tumor. PCR with primers spanning the breakpoints was able to detect mutant DNA molecules present at a level lower than 0.001% and identified mutated circulating DNA in patient plasma samples. These results highlight the sensitivity and specificity of the approach and suggest broad clinical utility of the PARE method. However, the approach has some limitations related to the loss of some rearranged sequences during tumor progression and the current high sequence cost.

Although the cost of genomic information has fallen steeply, the clinical translation of genetic risk estimates is not straight forward. Since the present analytical methods are not sufficient to make genetic data accessible in a clinical context, the clinical usefulness of these data for individual patients has not been formally assessed. A pioneer study undertook an integrated analysis of a complete human genome in a clinical context (Ashley et al. 2010). The genome of a patient with a family history of vascular disease and early sudden death had been sequenced, and 2.6 million SNPs and 752 CNVs showed increased genetic risk for myocardial infarction, type 2 diabetes and some cancers. Rare variations were found in three genes which are clinically associated with sudden cardiac death--- *TMEM43*, *DSP* and *MYBP3*. A variant in *LPA* was consistent with a family history of coronary artery disease. A heterozygous null mutation in *CYP2C19* suggested possible clopidogrel resistance and variants in *CYP4F2* and *VKORC1* might have a low initial dosing requirement for warfarin. The patient had several variants that are associated with good response to statins and one variant suggested that he might need a raised dose to achieve a good response.

This study provided an approach to comprehensively analyze a human genome in a defined clinical context. The authors assessed whole genome genetic risk, focusing on variants in genes that are associated with Mendelian diseases, novel and rare variants across the genome, and variants of pharmacogenomic importance. Additionally, this study developed an approach to interrogate disease risk across several common polymorphisms. However, there are still important limitations to comprehensively integrate genetic information into clinical care. Such as, a comprehensive database of rare mutations is needed and a continually updated pipeline is necessary. As whole genome sequencing becomes increasingly widespread,

availability of genomic information will no longer be the limiting factor in the application of genetics to clinical medicine. Development of methods integrating genetic and clinical data will assist clinical decision making and represent a large step towards individualized medicine. The transition to a new era of genome-informed medical care will need a team approach incorporating medical and genetics professionals, ethicists and health-care delivery organizations.

Two recent studies demonstrated the power that these whole genome sequence data hold for patients with a diagnosis of cancer. In the first report, Link and colleagues (Link et al. 2011) performed whole genome sequencing on the skin and bone marrow DNA of a patient with early-onset breast and ovarian cancer (negative for *BRCA1* and *BRCA2* mutations) and therapy-related acute myeloid leukemia (t-AML). The result revealed a novel and heterozygous 3 kb deletion removing three exons (7-9) of *TP53* in the normal skin DNA, which was homozygous deleted in the leukemia DNA due to uniparental disomy. Without whole-genome sequencing, this novel mutation in *TP53* would not have been discovered. Although the life of the patient could not be saved by the discovery, the implications for her children who may have inherited this mutation are immediate.

In the second study, Welch and colleagues (Welch et al. 2011) sequenced the DNA extracted from the leukemic bone marrow and a skin biopsy from a 39-year-old woman whose clinical presentation was consistent with acute promyelocytic leukemia, however, cytogenetic analysis revealed a different subtype associated with a poor prognosis. A novel insertional translocation on chr17 which created a pathogenic fusion gene *PML-RARA* was identified and this type of genetic event could not have been identified by traditional cytogenetic techniques. The whole genome sequencing results led to a change in AML therapy including giving the patient retinoic acid

which significantly improves the overall prognosis of patients with AML. Bone marrow transplantation is no longer considered in first remission. The remarkable achievement of this study is to generate comprehensive genomic data in a time frame that is clinically relevant for a patient.

With the fast development of new DNA sequencing technologies which are reducing the cost and accelerating our ability to study the complete genomic landscape, the somatic genetic changes that underline the initiation and progression of human cancer are rapidly becoming known. Cataloguing these changes and the frequency at which they occur in specific tumors and tumor subtypes is the first critical step on the path to understand what drives the oncogenic process. The complexity of the cancer genomes has been confirmed by the results from many cancer genome sequencing efforts to date and these achievements also solidify the concept that no two tumors are created equally even if they belong to the same histological subtype. Another impetus for studies of somatic genome alterations is the potential for therapies targeted against the products of these alterations. For example, treatment with the inhibitors of the epidermal growth factor receptor kinase (EGFR), gefitinib and erlotinib, leads to a significant survival benefit in patients with lung cancer who carry *EGFR* mutations, but no benefit in patients who carry wild-type *EGFR*. Therefore, comprehensive genome-based diagnosis of cancer is becoming increasingly crucial for therapeutic decisions.

Chapter Two: Long Span (DNA-PET) Sequencing Strategy for the Interrogation of Cancer Genomic Structural Variations

Introduction

The relatively short reads (~35-450bps) generated by current next generation high – throughput DNA sequencing technologies cannot allow us to sequence a genome in its entirety using de novo assembly, which means construct the genome sequence based solely on the sequencing reads without any other prior knowledge. The determination of a new genome sequence relative to a reference genome is often referred to as resquencing. In each resequencing project, there are five parameters that can be used (1) SNPs, (2) small indels (2-1000 bp), (3) large structural variations (>1000 bps), (4) new sequences which are not present in the reference genomes and (5) Genotype/haplotype information.

In general, the identification of SNPs and small indels is relatively straight forward although low sequence complexity is a confounding factor. However, the identification of large SVs is a routine procedure yet, because the extent of altered sequences at the SVs breakpoint and the presence of repetitive sequences can make SV detection difficult. The sequence reads cannot be unambiguously mapped the reference when the SVs reside in repetitive regions of the genome. Human DNA sequences could be broadly classified as four kinds of repeats: 1) segmental duplications also known as low copy repeats (LCR) (> 1 kb in length with 90% of sequence identity between copies), which are induced by recombinational processes and comprise about 5% of the human genome; the majority is located in the pericentromeric and subtelomeric regions. LCRs longer than 10kb and of over ~97% sequence identity can mediate or stimulate CNV formation by creating local genomic instability. LCRs have been shown to stimulate and/or mediate constitutional (i.e.,

inherited; both recurrent and nonrecurrent) evolutionary, and somatic genomic rearrangement (Gu et al. 2008); 2) simple sequence repeats (1-13 nucleotides) that arose by DNA polymerase slippage during replication; simple sequence repeats make up ~3% of the human genome; 3) transposable elements also known as interspersed repeats (few kilobases) which comprise the largest component of mammalian genomes (Babushok, D.V, 2006); 4) tandem repeats of the centromeric and telomeric regions which are often excluded from genome assemblies. In addition, SVs are often complex, with the occurrence of multiple events in close proximity, so that the events cannot be correctly detected by short reads. Although SVs account for a large proportion of base pairs which are affected by variation in the genome (Kidd et al. 2008; Korbel et al. 2007), it is likely that a substantial number of SVs are missed in most genome sequencing projects (Table 2.1) because of current limitations in the extraction of SV information from short read genome sequencing data.

The PET technologies differ in the length of the DNA fragments that are sequenced from both ends. Currently, the commonly used fragment lengths for paired-end sequencing range from 200 bp to 3 kb. In theory, small insert size (sub-kilo base) has the advantages of a tight size range of DNA fragments and thereby a greater sensitivity for the detection of small insertions and deletions. In contrast, large insert size (one kilo base to tens of kilo bases) has the advantage of higher genomic physical coverage with the drawback of less precise localization of breakpoint regions. It is expected that a combination of paired-end reads from different length fragments will provide optimal SV detections. To address this question, our group developed a robust procedure to construct larger insert size libraries (10-20 kb) and we used three well established cancer cell lines including the breast cancer cell line MCF-7, the colon cancer cell line HCT116 and the chronic myelogenous leukaemia (CML) cell

line K562 as test genomes. We compared the characteristics of the large insert size libraries (10 kb and 20 kb) with short insert size (1 kb) libraries at the same sequencing depths and costs. Although short insert size libraries bear an advantage in identifying small deletions, they do not provide a significantly better breakpoint resolution. Large insert size libraries are superior to short insert size libraries in providing higher physical genome coverage and therefore achieve greater sensitivity for the identification of the different types of SVs, including copy number neutral and complex events. Further, large insert size libraries allow the identification of SVs within repetitive sequences which cannot be spanned by short inserts.

Table 2.1: Individual genomes which have been sequenced by PET technology

Project	Technology	SVs	Fragment size	Reference
African, European	Illumina	5,740	200bp	Bentley 2008
Lung cancer genome	Illumina	409	500bp&200bp	Campbell et al. 2008
AML genome	Illumina	726 (1-30bp)	150-200bp	Ley et al. 2008
AML genome	Illumina	Limited	95bp to 298bp	Mardis et al. 2009
24 human breast cancer	Illumina	2,166	500bp	Stephens et al. 2009
HapMap sample	SOLiD	5500 (unknown definition)	1.4kb	McKernan et al. 2009
Melanoma genome	Illumina	51	200bp & 400bp	Pleasance et al. 2010a
Lung cancer genome	SOLiD	392	500-600bp & 3-4kb	Pleasance et al. 2010b
Giloma cell line	SOLiD	1314	1.4kb	Clark et al. 2010
A patient	Heliscope	752 CNVs	100-500bp	Ashley, 2010
Four colorectal and two breast cancers	SOLiD	54	1.4kb	Leary et al. 2010
A child with bilateral, young-onset Wilms tumor	Illumina	1	3kb	Slade et al. 2010
A Japanese individual	Illumina	5,488	200bp	Fujimoto et al. 2010
A basal like breast cancer	Illumina	302	200bp	Li Ding et al. 2010
A lung cancer tumor	Complete Genomics	344	200-400bp	Lee et al. 2010
13 pancreas cancer	Illumina	558	400-500bp	Campbell et al. 2010
A hepatocellular carcinoma	Illumina	33	250bp	Totoki et al. 2011
Seven primary human prostate cancers	Illumina	755	400bp	Berger et al. 2011
CML and bone cancer	Illumina	371	400-500bp	Stephens et al. 2011
Multiple myeloma	Illumina	475	370-410bp	Champman et al. 2011

Results

DNA-PET library construction and sequencing

We randomly sheared the genomic DNA from MCF-7, HCT116, and K562 to the desired size range and gel-purified a narrowed range of 1 kb, 10 kb, and 20 kb DNA fragments, respectively. The DNA fragments were subjected to DNA-PET library construction and sequencing to generate PET sequences with each tag of 25 bp tag length (details in Materials and Methods). The PET sequences (2 x 25 bp) were mapped to the reference genome (NCBI build 36) with SOLiD System Analysis Pipeline Tool, Corona Lite, allowing 2 color-code mismatches (Figure 2.1). In total, we generated seven genomic DNA-PET datasets, two each for MCF-7 and HCT116 (1 kb and 10 kb) and three for K562 (1 kb, 10 kb and 20 kb, Table 2.2).

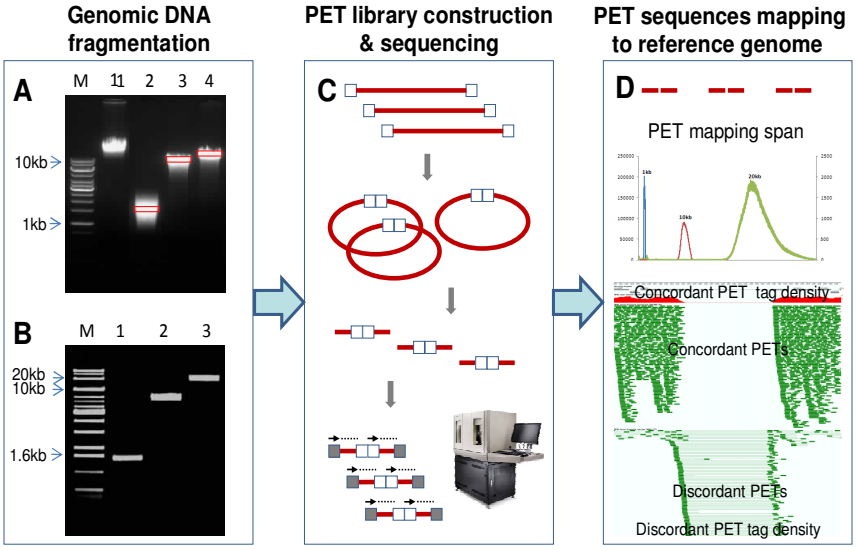

Figure 2.1. DNA-PET library construction, sequencing and mapping.

(A) The genomic DNA was randomly sheared to different size range. (B) The very narrow region DNA fragments were obtained after size selection. (C) The purified DNA fragments were circularized, *EcoP15I* digested, sequencing adaptor ligated and finally sequenced by SOLiD sequencer. (D) PET mapping span distribution in 1kb (blue), 10kb (red) and 20kb (green) libraries. Based on the mapping pattern, PET can be distinguished as concordant PET and discordant PET.

To get comparable non-redundant PET numbers among different insert size libraries of the same genome, we randomly selected a subset of original MCF-7 and HCT116 10 kb libraries (Hillmer et al. 2011) resulting in approximately 20 million non-redundant (NR) PET sequences for all MCF-7 and HCT116 libraries. The 18.4 million non-redundant and uniquely-mapped PET sequences of the MCF-7 10 kb library resulted in approximately 69-fold physical (fragment) coverage of the human genome. Of these PET sequences, 88% (16.3 million) were mapped concordantly to the reference genome as concordant PETs (cPETs) (5' and 3' tags of a PET mapped on the same chromosome, same strand in 5'→3' orientation and within the expected insert size). The MCF-7 10 kb cPETs were mapped within the range of 8,099 bp to 16,217 bp and with a peak of the insert size distribution at 11,273 bp. This proportion of PETs reflected the agreement of the genome architecture between MCF-7 and the reference genome. In contrast, 11.6% (2.1 million) of the uniquely-mapped PETs were mapped as discordant PETs (dPETs) (5' and 3' tags of a PET mapped too far away or too close to each other, or mapped in reversed order as 3'→5', or mapped on different strands or different chromosomes). These dPETs were indicative of potentially rearranged genomic regions crossing the breakage/fusion junction points where the MCF-7 genome is different from the reference genome. In the MCF-7 1 kb dataset, 96% (17,598,541 of 18,335,127) NR PET were concordantly mapped to the reference genome and 4% (736,586) were discordant. The lower proportion of dPETs in the 1 kb library is due to the low number of short span PET constructs which comprise the largest proportion of dPETs for 10 kb libraries (Hillmer et al. 2011). As expected, the physical coverage of the MCF-7 1 kb library was considerably lower (8-fold) compared to the 10 kb library (69-fold) with the same number of NR PETs (18 million). In K562, the comparable number of NR PETs (~ 40 million)

of the 1 kb, 10 kb and 20 kb libraries gave physical coverage of 15-fold, 122-fold, and 277-fold, respectively.

Table 2.2. DNA-PET library statistics

	MCF7		HCT116		K562		
	IHM072 (1kb)	IHM001 (10kb)	IHH021 (1kb)	IHH001 (10kb)	IHK002004 (1kb)	IHK006007 (10kb)	IHK016017 (20kb)
Total Tags	270,334,239	233,715,372	234,045,521	357,273,623	311,788,203	564,115,773	410,069,683
Mappable Tags	73,484,523	97,200,820	106,838,016	212,486,672	183,080,988	335,776,419	246,267,355
PET	20,899,040	22,856,160	28,389,915	53,175,650	85,086,071	133,675,193	83,261,984
NR PET [1)]	18,335,127	18,432,387	20,613,096	20,610,717	41,996,278	44,065,447	38,393,242
Redundancy	1.14	1.24	1.38	2.58	2.03	3.03	2.17
Span Range	1,155-1,514	8,099-16,217	1,186-1,574	7,200-11,780	880-1,320	6,846-10,248	15,590-33,140
Median	1,347	11,273	1,367	8,514	1,100	8,303	21,700
Coverage	8.2	69.3	9.4	58.5	15.4	122.0	277.7
cPET [2)]	17,598,541	16,292,711	18,919,833	17,938,263	39,818,591	40,261,305	35,246,258
cPET % [3)]	96.0%	88.4%	91.8%	87.0%	94.8%	91.4%	91.8%
dPET [4)]	736,586	2,139,676	1,693,263	2,672,454	2,177,687	3,804,142	3,146,984
dPET % [5)]	4.0%	11.6%	8.2%	13.0%	5.2%	8.6%	8.2%
Singleton	728,445	2,119,094	1,688,547	2,661,375	2,157,688	3,771,859	3,103,814
Singleton % [6)]	98.9%	99.0%	99.7%	99.6%	99.1%	99.2%	98.6%
dPET cluster (≥2)	1,689	1,663	997	1,422	4,432	3,225	7,636
dPET [7)]	8,141	20,582	4,716	11,079	19,999	32,283	43,170
% [8)]	1.1%	1.0%	0.3%	0.4%	0.9%	0.8%	1.4%

1) non-redundant PET
2) concordant PET
3) Proportion of concordant PET to NR PET
4) discordant PET
5) Proportion of discordant PET to NR PET
6) Proportion of singleton to dPET
7) dPET which belong to the cluster size ≥2
8) Proportion of dPET which belong to the cluster size ≥2 to total dPET

Large structural variations identified by different insert size libraries

Each of the dPETs could potentially map over a breakage/fusion point. However, it is inevitable that spurious dPET mappings would arise due to chimeric ligation during the construction of DNA-PET libraries or incorrect tag mapping. To reduce such random noise, we used the PET mapping overlap scheme (clustering, detail in Materials and Methods) to discard all the singleton and cluster size 2 dPETs and considered only the dPET clusters with multiple overlapping PETs (≥3) as true signals for further analysis of genome rearrangements (Table 2.3; Figure. 2.2).

Table 2.3. dPET cluster statistics

Library	Total	Singleton	Cluster 2	Cluster 3	Cluster 4	Cluster 5	cluster >5
MCF7 1kb	736,586	728,445	1,330	1,176	928	625	4,082
%	100%	98.89%	0.18%	0.16%	0.13%	0.09%	0.55%
MCF7 10kb	2,139,676	2,119,094	1,268	459	388	385	18,082
%	100%	99.04%	0.06%	0.02%	0.02%	0.02%	0.85%
HCT116 1kb	1,693,263	1,688,547	850	555	488	385	2,438
%	100%	99.72%	0.05%	0.03%	0.03%	0.02%	0.14%
HCT116 10kb	2,672,454	2,661,375	1,330	462	320	285	8,682
%	100%	99.59%	0.05%	0.02%	0.01%	0.01%	0.32%
K562 1kb	2,177,687	2,157,688	5,430	1,560	1,236	1,175	10,598
%	100%	99.08%	0.25%	0.07%	0.06%	0.05%	0.49%
K562 10kb	3,804,142	3,771,859	3,682	894	636	495	26,576
%	100%	99.15%	0.10%	0.02%	0.02%	0.01%	0.70%
K562 20kb	3,146,984	3,103,814	10,888	3,048	1,504	820	26,910
%	100%	98.63%	0.35%	0.10%	0.05%	0.03%	0.86%

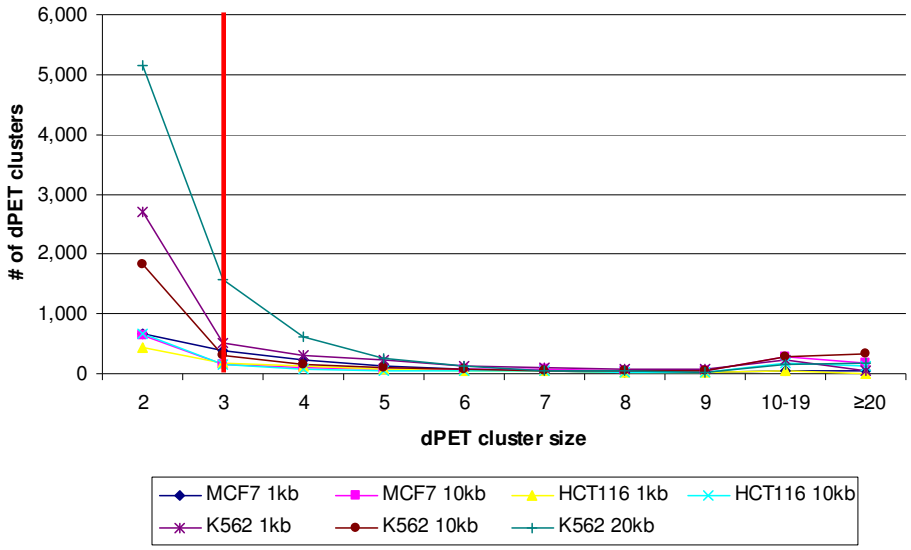

Figure 2.2. dPET cluster size distribution.

The number of dPET clusters (y-axis) is shown for the individual cluster sizes (x-axis). Red vertical line represents the cutoff for dPET clusters regarded as reliable breakpoint pairs (size three and higher).

The large insert size and high sequencing depth of the libraries allowed for the identification of different types of SVs, including deletions (the 5' mapping anchor region was far apart from the 3' mapping anchor region), tandem duplications (the mapping order was 3' to 5' instead of the normal 5' to 3'), unpaired inversions (the mapping orientation was reversed, on different strand), isolated translocations (the 5' and 3' anchors mapped to different chromosomes). Two closely positioned dPET clusters could be used to deduce the SVs with two rearrangement points such as inversions, insertions, and balanced translocations (details in Materials and Methods). In addition and in contrast to previous studies, eight different sub-types of insertions were characterized in this study. The eight sub-types of insertions included different combinations of intra or inter-chromosomal, direct or inverted, forward or backward insertions (Figure 2.3).

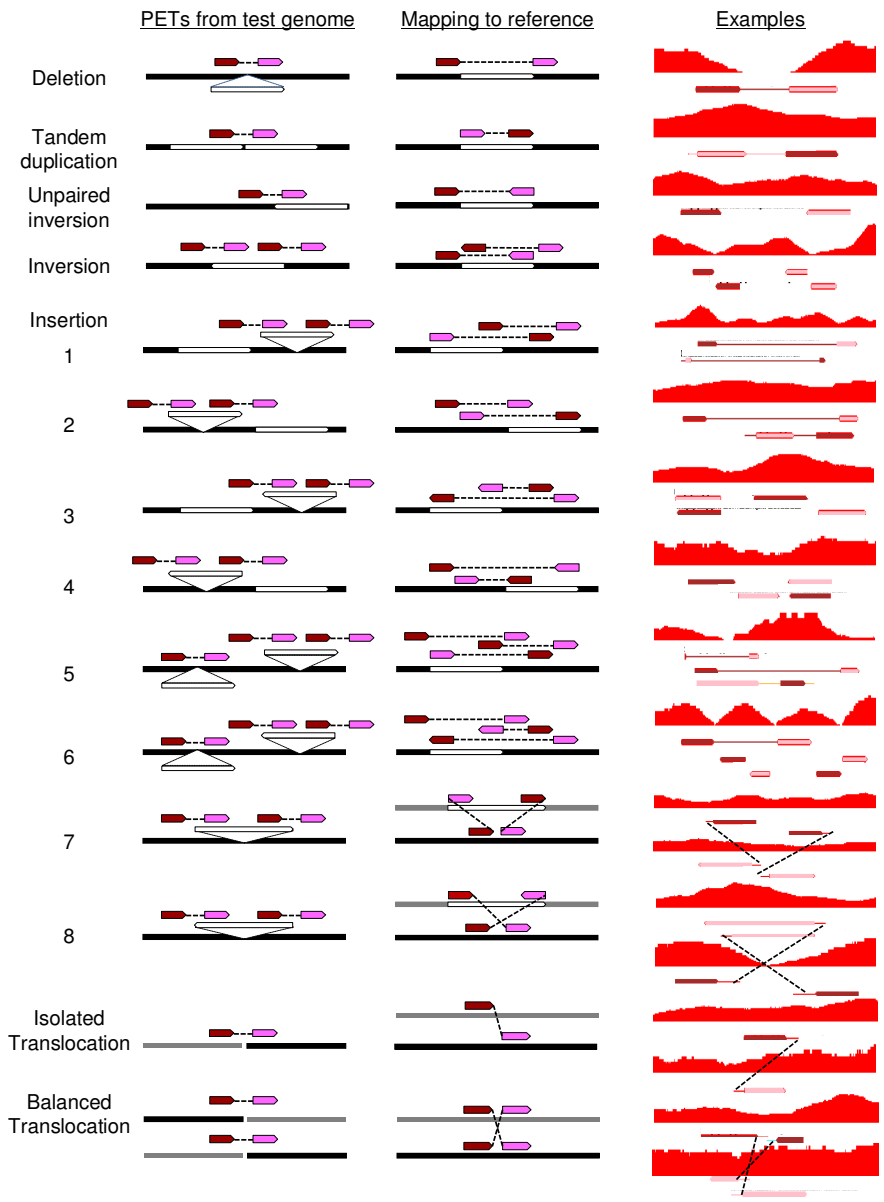

Figure 2.3. SVs identification based on the mapping pattern of dPET clusters.

The dark red and pink arrows represent the 5' and 3' anchor region of the dPET cluster. Black, white and gray horizontal lines represent chromosome segments. The sub-type of insertion as follows: (1) Intra chromosome direct forward insertion. (2) Intra chromosome direct backward insertion. (3) Intra chromosome inverted forward insertion. (4) Intra chromosome inverted backward insertion. (5) Deletion plus intra chromosome direct forward insertion. (6) Deletion plus intra chromosome inverted forward insertion. (7) Inter chromosome direct insertion. (8) Inter chromosome inverted insertion.

A high count for a dPET cluster (large number of dPETs spanning the same breakage/fusion point) gives high confidence for the rearrangement point and may also reflects the copy number of the breakage/fusion point. The highest dPET cluster count in the MCF-7 10kb library was 766 and only 91 in the corresponding 1 kb library. We observed similar drops in cluster count for HCT116 (148 in the 10 kb library and 63 in the 1 kb library) and K562 (2,106 for the 20 kb library, 692 for the 10 kb library and 127 for the 1 kb library, respectively). Using cluster count 3 as cut off, the numbers of identified SVs in the 10 kb library of MCF-7, HC116, and K562 were 899, 654, and 1,570 (Table 2.4). The total number of SVs identified in the 1 kb libraries of each genome was comparable to the number of SVs found in the 10 kb libraries; however, the composition of the SV types was different. In the 1 kb libraries, the vast majority of SVs was deletion (79% in MCF-7, 80% in HCT116, and 78% in K562); whereas in the 10 kb libraries, the percentage of deletions was much lower (33% in MCF-7, 59% in HCT116, and 38% in K562). In contrast, the number of inversions and insertions identified in 1 kb libraries was much lower than in 10 kb libraries (Table 2.4 and Table 2.5).

Table 2.4. SVs in individual library

Sample	Library size	Deletion	Tandem Duplication	Inversion	Insertion		Unpaired inversion	Isolated translocation	Balanced translocation
					Intra chr	Inter chr			
MCF7	1kb	739	51	1	1	4	46	93	0
	10kb	300	226	30	13	23	126	180	1
HCT116	1kb	415	32	4	1	4	29	35	0
	10kb	385	86	33	13	13	55	69	0
K562	1kb	1,252	110	14	5	9	85	123	0
	10kb	590	203	35	39	16	205	482	0
	20kb	980 (3) 211 (5)*	106	47	19	15	104	162	0

*The large span window of this library created many artificial deletion clusters; hence we increased the cluster size cut off from 3 to 5 to reduce the number of false positive calls.

Table 2.5. Sub-types of insertions in individual libraries

Sub-type of insertion	MCF7		HCT116		K562		
	1kb	10kb	1kb	10kb	1kb	10kb	20kb
Intra chromosome forward direct insertion	0	4	1	7	3	12	2
Intra chromosome backward direct insertion	0	0	0	1	2	7	8
Intra chromosome forward inverted insertion	1	5	0	2	0	10	2
Intra chromosome backward inverted insertion	0	3	0	2	0	8	6
Deletion plus intra chromosome forward direct insertion	0	0	0	0	0	1	0
Deletion plus intra chromosome forward inverted insertion	0	1	0	1	0	1	1
Inter chromosome direct insertion	0	8	2	8	4	10	9
Inter chromosome inverted insertion	4	15	2	5	5	6	6
Total	5	36	5	26	14	55	34

Classify isolated and complex SVs in each library

In some amplified loci of cancer genomes, due to the complicated genomic architecture, a large number of dPET clusters were connected to form complex rearrangement units. In these units, it might be misleading to assign a particular SV type to a dPET cluster e.g. if the breakpoints of a tandem duplication are surrounded by deletions and/or translocations, the rearrangement might not be interpreted as a tandem duplication. Therefore, a breakpoint based interconnection network was established to separate breakpoints in complex regions from isolated and less complex SVs. (Figure 2.4 and details in Materials and Methods). Non-complex deletions, tandem duplications, unpaired inversions, and isolated translocations involve only one cluster, which was reflected by a supercluster count 1. Superclusters with count 2 comprised inversions, insertions and balanced translocations which involved two dPET clusters. Deletions or tandem duplications plus insertions or inversions were the major source of superclusters with count 3. We observed a significant number of SVs that were connected by multiple dPET clusters to form complex rearrangement units in which the exact architecture was complicated by overlapping of multiple SV events. Therefore, we classified supercluster count 4 or more connected SVs as complex. The supercluster analysis result showed that 1 kb library contained more isolated SV events than 10 kb libraries and the major contribution came from deletions (Figure 2.5). MCF-7 had the most complicated rearrangement unit and higher supercluster counts than HCT116 and K562. The largest rearrangement unit in the MCF-7 10 kb library involved 212 dPET clusters, of which 210 were located in the amplified regions on chromosomes 1, 3, 17, and 20. The highest supercluster count of the 10 kb library in K562 and HCT116 was 115 and 41, respectively. The lower highest supercluster count in HCT116 indicated less rearrangement of this genome compared to MCF-7 and K562.

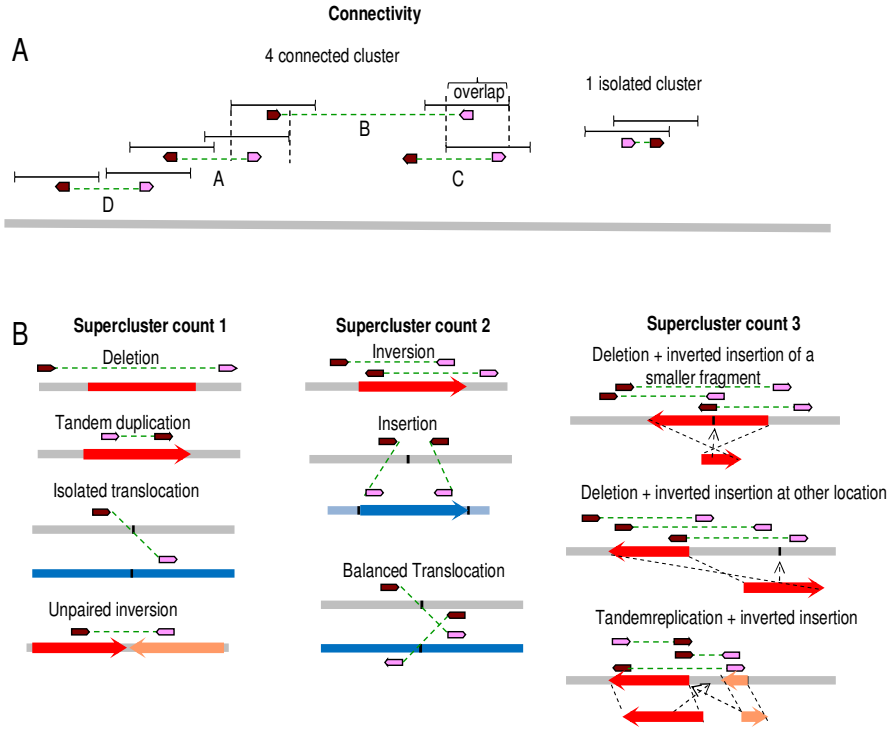

Figure 2.4. Supercluster definition and categories.

(A) Connectivity of breakpoints. An interconnection network was established by grouping together dPET clusters (to superclusters) which have at least one anchor region within 10kb of each other. This may result in an indirect connection of cluster D and C. 5' anchor regions are indicated in dark red, 3' anchor regions are indicated in pink. (B) Examples of single, double and triple supercluster.

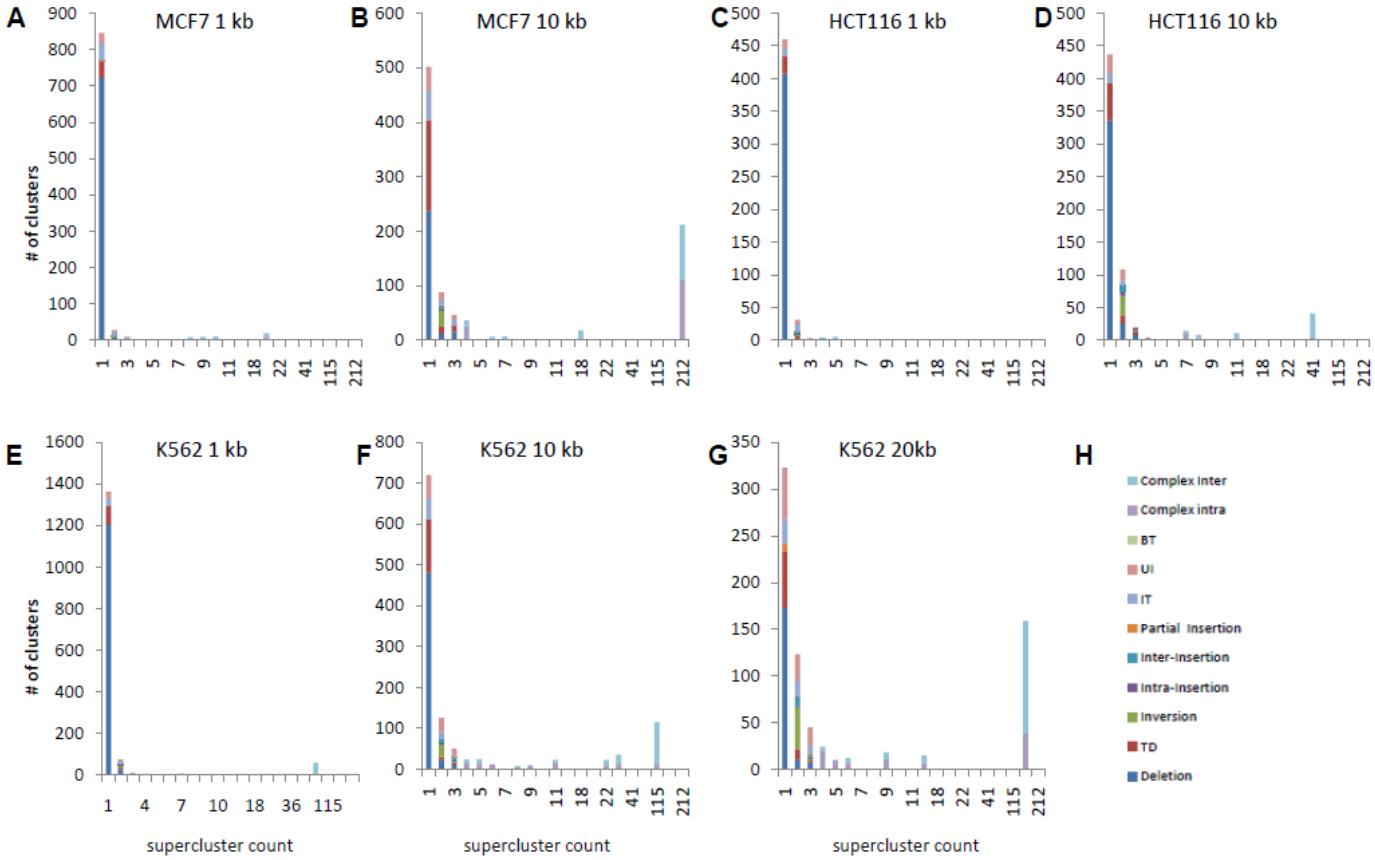

Figure 2.5. Supercluster statistics in individual library.

(A-G) Distribution of degrees of connectivity represented by superclusters in each library. Numbers of clusters (y-axis) for each supercluster count (number of interconnected clusters, x-axis) is shown. (H) Color code of each kind of SV.

Comparison of isolated SVs from 1 kb, 10 kb and 20 kb DNA-PET libraries

To determine the sensitivity of different insert sizes to identify SVs, we compared SVs from different insert size libraries of each genome. We excluded complex SVs from this analysis but included clusters of size 2 for SVs which matched an SV in another library. For 10 kb and 20 kb libraries specific SVs, we increased cluster count cut off from 3 to 6 to increase the confidence. The result showed that 1 kb libraries could identify more deletions than 10 kb libraries. Such 1 kb library-specific deletions were usually smaller than 5 kb. The span of most deletions detected by 10 kb libraries ranged between 5 kb and 500 kb (Figures 2.6.A).

Inversions are prone to map to segmental duplications (Stefansson et al. 2005). The ambiguous mapping for sequence tags of segmental duplications results in the exclusion of such tags in most pipelines and the inability to identify breakage/fusion points in these regions. Small insert libraries such as 1 kb libraries are less likely to span such regions of ambiguous mapping and are expected to have lower detection rates for inversions. The lower physical coverage of 1 kb libraries, as compared to 10 kb libraries further limits their ability in identifying inversions. The MCF-7 1 kb library resulted in only 1 inversion with a cluster size ≥3 (Table 2.4). However, the comparison with the 30 inversions identified by the 10 kb library indicated that 3 inversions matched with at least one low confidence cluster of size 2 in the 1 kb library. Similarly, we identified 34 inversions by the 10 kb library in HCT116 and found evidence by low confidence clusters for only 4 inversions by the respective 1 kb library. Compared to 1 kb libraries of MCF-7 and HCT116, the higher total PET number of the K562 1 kb library resulted in a higher physical coverage and the identification of more inversions. In K562, there were two 1 kb library specific inversions with a span of 509 bp and 836 bp, respectively. The 10 kb library data of K562 contained one of the two clusters which identify an inversion.

The difference in the detection rate of insertions between 1 kb and 10 kb libraries was comparable to inversions. There was no 1 kb library-specific insertion in MCF-7. In HCT116 and K562, two and one 1 kb library-specific insertions were identified, respectively, and all were smaller than 1 kb. Of these three small insertions, the respective 10 kb libraries identified one of the two breakage/fusion points. In summary, compared to 1 kb libraries,
10 kb libraries had a higher sensitivity in identifying more and large span SVs. All the SVs which were specific to 1 kb libraries and which were missed by 10 kb libraries had either a low cluster count or a short span (Figure. 2.6).

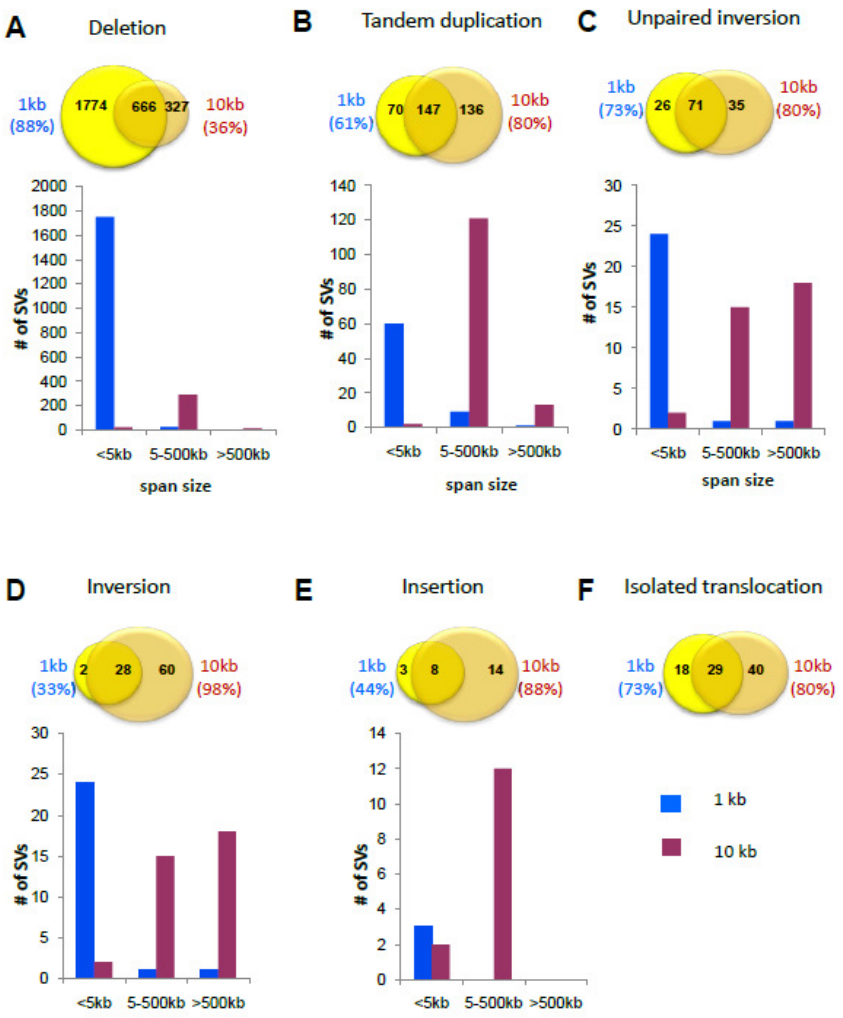

Figure 2.6. Comparison of number and span distribution of specific SVs identified by 1 kb and 10 kb libraries in the three genomes.

Venn diagrams showing the respective numbers of SVs in each library type and the overlap of SVs. Number of SVs (*y-axis*) of the indicated SV categories (A-D) were shown for the different span sizes (*x-axis*).

The comparison of the ability to identify SVs by the 10 kb and 20 kb libraries of K562 showed a slightly higher detection rate of inversions and unpaired inversions for the 20 kb library and a higher detection rate of deletions for the 10 kb library (Table 2. 4). The span of the majority of the events from these two libraries was comparable, between 5 kb to 500 kb (Figure.2.7). One hundred and ninety-four deletions were identified by the 10 kb but not the 20 kb library. The vast majority (188) were short deletions (< 20 kb). Sixteen deletions were specific to the 20 kb library and half of them had either both or one anchor region located in a repetitive region which could not be spanned by the 10 kb library (Figure 2.8). Similarly, tandem duplications which were exclusively identified by the 10 kb library were small in size (<20 kb). The 20 kb library-specific unpaired inversions or inversions which were missed by the 10 kb library showed a low cluster count or had one/both anchors located in repetitive regions.

After the comparison of the SVs from different insert size libraries, we combined the common SVs from different insert size libraries and obtained the genome-specific structural variations (Table 2.6, Appendix Table 1-3). In these three genomes, deletion and tandem duplication were the most abundant SVs whereas the number of inversion and insertion was less than other type of SVs. HCT116 showed the lower number of complex SVs, suggesting a lower degree of rearrangements compared to MCF-7 and K562.

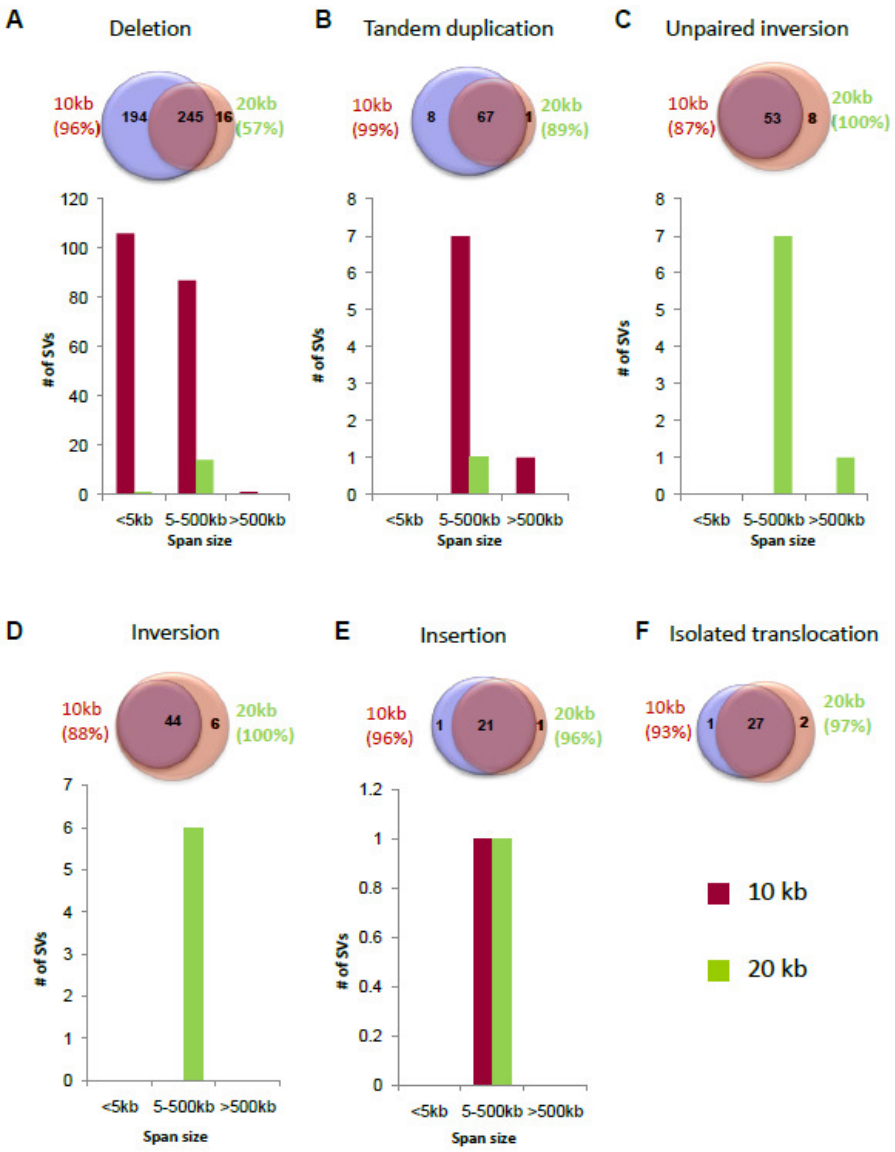

Figure 2.7. Comparison of number and span distribution of specific SVs identified by 10 kb and 20 kb libraries in K562.

Venn diagrams showing the respective numbers of SVs in each library type and the overlap of SVs. Number of SVs (*y-axis*) of the indicated SV categories (A-E) were shown for the different span sizes (*x-axis*).

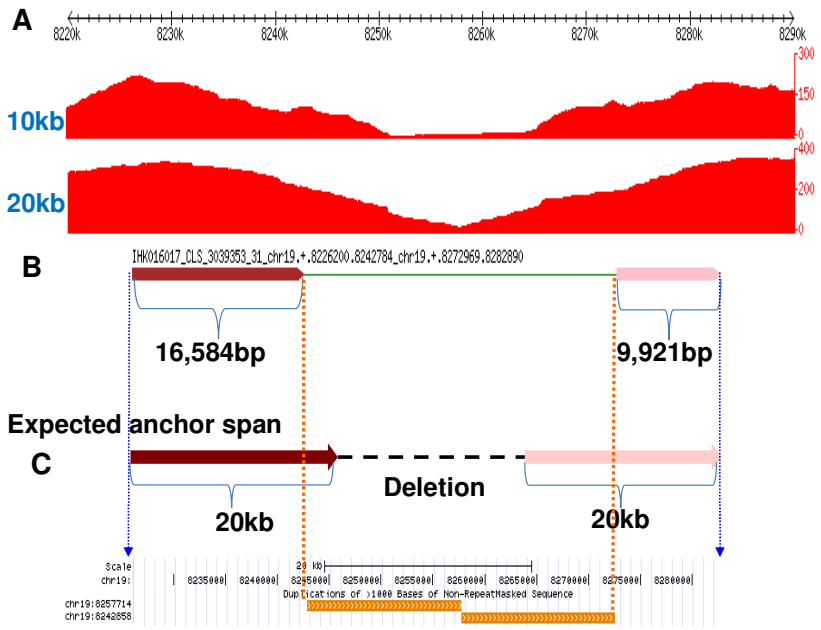

Figure 2.8. Example of a 20 kb library specific deletion in K562.

(A) The drop of expected coverage (red track) in 10 kb and 20 kb library indicates the presence of a deletion. (B) The deletion only could be detected by a 20 kb library dPET cluster (cluster size 31) and the 5' and 3' anchor span were 16,584 bp and 9,921bp. The red and pink arrows represented the 5' and 3' anchor regions of the dPET cluster. (C) The expected anchor span of 20 kb library is 20 kb and the segmental duplications (orange blocks) located at this deletion created the shorter anchor span in 20 kb library. The PETs of 10 kb library could not cross the segment duplications resulting in the failure to detect this deletion by the 10 kb library.

Table 2.6. SVs identified in each genome

	Isolated							Complex	
	Deletion	TD [1)]	Inversion	Intra-chr Insertion	Inter-chr Insertion	UI [2)]	IT [3)]	Intra-chr	Inter-chr
MCF7	845	153	23	4	2	50	49	144	172
HCT116	569	71	27	4	4	35	16	28	50
K562	1393	145	52	16	9	83	38	247	260

1) tandem duplication
2) unpaired inversion
3) isolated translocation

Validation of predicted SVs and Resolution of 1kb and 10kb libraries

Genomic PCR and consequent Sanger sequencing were used to confirm the breakpoints of 161 randomly selected SVs of the three genomes and a total of 129 SVs (80%) were confirmed (Table 2.7). Twenty of the 32 SVs which could not be validated had a cluster count <10 indicating low cluster count represented a lower confidence of SV predication. Eight of the non-validated SVs (cluster size 11 to 50) had shorter anchor spans which suggested that repetitive sequences around the breakpoints inhibited the PCR amplification or mapping error. The remaining two unpaired inversions (cluster counts 156 and 122) and two isolated translocations (cluster counts 113 and 245) were within complex regions with high supercluster count. The intersection of different breakpoints most likely inhibited the PCR amplification.

Table 2.7. Genomic PCR and Sanger sequencing validation statistics

SVs	Investigated	Validated[1)]	Non-validated[2)]	Success%
Deletion	62	51	11	82
Tandem duplication	28	21	7	75
Inversion	14	13	1	93
Insertion	7	6	1	86
Unpaired inversion	16	10	6	63
Isolated translocation	34	28	6	82
Total	161	129	32	80

[1)] PCR products with single band at expected size range
[2)] No PCR products or PCR products with multiple bans

We calculated the breakpoint resolution of 1 kb and 10 kb libraries and defined the resolution as the genomic distance in bp between the dPET clusters predicted breakpoint coordinate and the breakpoint coordinate determined by genomic PCR and Sanger sequencing. Inversions were

excluded from the resolution calculation as they tend to be located in repetitive regions which do not allow the unambiguous positioning of the breakpoints. In total, 244 breakpoints were used to calculate the resolutions (242 in 10 kb libraries and 140 in 1 kb libraries) (Table 2.8). For both 10 kb and 1 kb library, the highest resolution was 0 bp and the lowest resolution for the 10 kb libraries was 10,799 bp and 1,205 bp for the 1 kb libraries. Importantly, the median resolution was 377 bp for 10 kb libraries and 115 bp for 1 kb libraries. This indicated that the higher coverage of the large insert libraries provides a resolution for the majority of breakpoints which is comparable to small insert libraries.

A large distance between the predicted breakpoints by dPET clusters and the true breakpoint locations is indicative of repetitive sequences which prevent a unique mapping. Of the 244 confirmed breakpoints, 104 were identified only by a 10 kb library, 138 were identified by 1 kb and 10 kb libraries, and 2 were identified only by a 1 kb library. Of all the 10 kb library specific breakpoints, 38/104 (36.5%) had a breakpoint resolution larger than 1.5 kb, whereas only 12/138 (8.7%) of 1 kb and 10 kb library common breakpoints had a resolution larger than 1.5 kb (Figure 2.9. A; $P < 10^{-9}$ [Fisher Exact Test]). Manual investigation of 10 kb specific breakpoints with resolutions larger than 1.5 kb showed that 33 of the 38 breakpoints (87%) had repetitive sequences covering the distances between predicted and true breakpoints, especially in the 1.5 kb regions next to the confirmed breakpoints (Figure 2.9. B). Only 4 of the 12 (33%) 1 kb and 10 kb common breakpoints were covered by repetitive sequences. The repetitive sequences were discontinuous and tags of the 1 kb dPET mapped to the gaps between the repetitive sequences allowing for the identification of the respective SVs (Figure 2.9. C). This strongly suggested that larger repetitive sequences around breakpoints prevent mapping of tags of 1 kb libraries and hence the identification of the respective SVs.

Table 2.8. Breakpoint resolution of 1 kb and 10 kb libraries

	Total		Deletion		Tandem duplication		Unpaired inversion		Isolated translocation		Insertion	
	1kb	10kb	1kb	10kb	1kb	10kb	1kb	10kb	1kb	10kb	1kb	10kb
Breakpoints	140	242	54	100	28	42	12	20	36	56	10	24
Highest (bp) [1)]	0	0	4	0	5	2	28	15	0	3	51	93
Lowest (bp) [2)]	1,205	10,799	1,067	5,788	878	6,930	613	1,407	1,123	10,799	1,205	9,668
Median (bp)[3)]	115.5	377	191.5	420.5	334	369.5	231	711	49	236.5	183	634

[1)] Highest: The smallest difference between the genomic PCR and Sanger sequence confirmed breakpoints and DNA-PET predicted breakpoints

[2)] Lowest: The largest difference between the genomic PCR and Sanger sequence confirmed breakpoints and DNA-PET predicted breakpoints

[3)] Median: The median difference between the genomic PCR and Sanger sequence confirmed breakpoints and DNA-PET predicted breakpoints

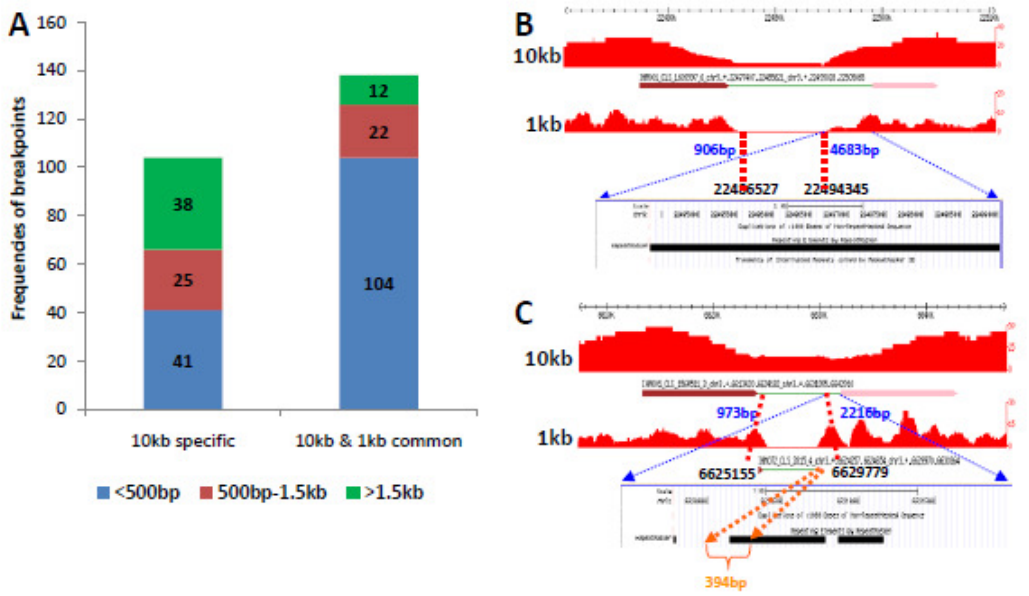

Figure 2.9. Breakpoint resolution and repetitive sequences of 1 kb and 10 kb libraries specific and common SVs

(A) Breakpoints confirmed by PCR and Sanger sequencing and their resolution (defined as the distance in bp between the predicted and actual breakpoints. (B) A 10 kb library specific deletion in MCF-7. The genomic PCR and Sanger sequencing confirmed left and right breakpoints are chr9:22,486,527 and chr9:22,494,345, respectively. The resolution of the left and right sides of the deletion are 906 bp and 4,683 bp, respectively. Repetitive sequence which does not allow unambiguous mapping covers the entire 4,683 bp region. The repetitive sequence could not be spanned by the 1 kb library preventing the identification of this deletion. (C) A deletion in MCF-7 identified by both 10 kb and 1 kb libraries. The left and right breakpoints confirmed by genomic PCR and Sanger sequencing are at chr3:6,625,155 and chr3:6,629,779, respectively. Based on the 10 kb library predicted breakpoints, the resolution on the left and right sides of the deletion are 973 bp and 2,216 bp, respectively. The tags of the 1 kb library mapped to the gap (in orange) between the repetitive sequences and allowed the identification of this deletion.

Copy number analysis by PET

Genomic rearrangements, such as deletion, tandem duplication, insertion and isolated translocation, give rise to copy number variations (CNVs) in the cancer genome (Shaikh et al. 2009). CNV is one of the key genetic drivers of diseases such as cancer (Beroukhim et al. 2010). Karyotyping, FISH and aCGH have been extensively used in genome-wide CNV prediction. More recently, the high-throughput sequencing for whole genome at high coverage has motivated the mapping of genome-wide CNVs using sequenced reads mapped onto the reference genome, based on the fact that the number of reads mapped to a genomic region is expected to be proportional to its copy number (Campbell et al. 2008; Xie et al. 2009). The copy number estimates and genomic rearrangements acquired from paired-end sequencing data have revealed the relationship between genomic breakpoints and amplicons or deletions (Pleasance et al. 2010a; Pleasance et al. 2010b; Stephens et al. 2009).

We analyzed the copy number changes across the genome of MCF-7, HCT116 and K562 by computing the density of all cPETs from 10 kb libraries. By comparison of the GC content from different size libraries, we confirmed that the GC bias in each library was due to the sample preparation step, but not from sequencing platform. Then the GC bias was corrected by calculating the GC content of the corresponding sliding windows along the genome. Another advantage of using cPET tags was to reduce the PCR amplification error by keeping one of the PETs which were mapped to the exact locations. The result showed that four significant amplified regions of MCF-7 are located on chromosome 1, 3, 17 and 20 (Figure 2.10); three amplified regions in K563 genome are located on chromosome 9, 13 and 22 whereas there is no significant amplified region could be found in HCT116 genome. The copy number prediction of MCF-7 was compared with the copy number estimation for MCF-7 by a

244K array comparative genomic hybridization (aCGH) as a validation and the result showed a correlation of r2=0.776 (Figure. 2.11). In addition, copy number estimation using sequencing data produced more and smaller copy number segments, 257 by aCGH and 714 based on cPET tags.

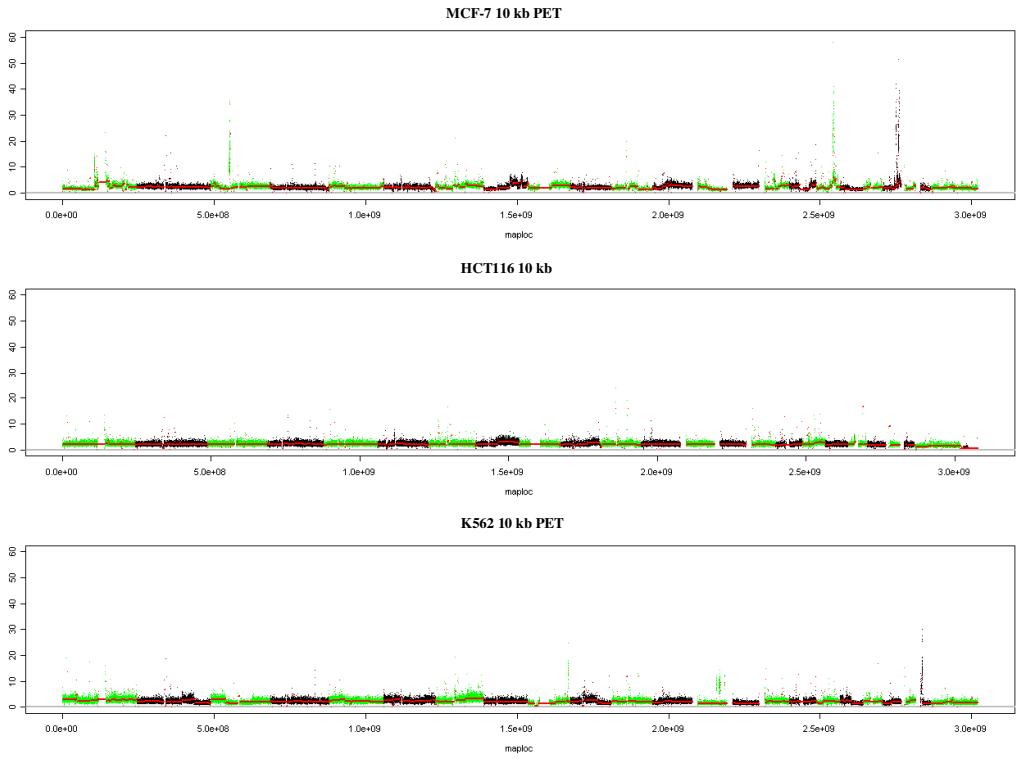

Figure 2.10. Whole genome copy number of the three genomes estimated by cPET tags of 10 kb libraries.

The green and black dots represent copy number of the probes or windows of each chromosome, and the red lines represent the copy number segments after smoothing.

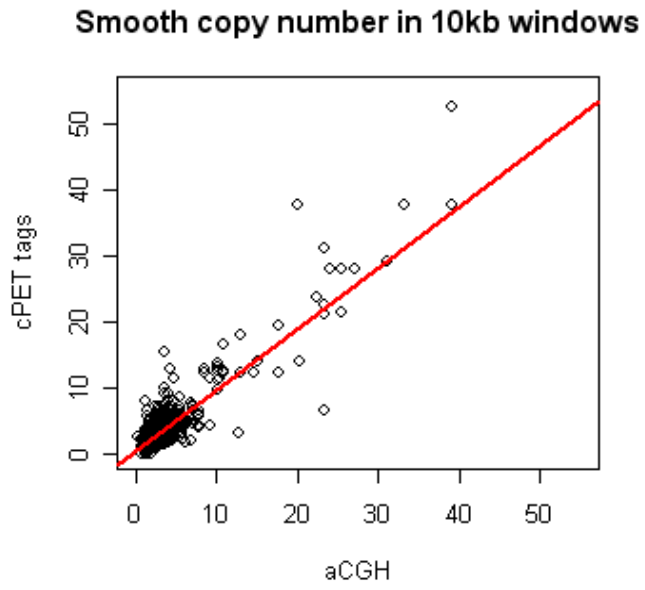

Figure 2.11. Correlation of copy number estimation for MCF-7 by aCGH and high throughput PET sequencing.

The estimated copy number of genomic segments of at least two adjacent windows with the same copy number estimate (y-axis) were correlated with copy number estimates by aCGH (x-axis): r2=0.7763.

A large homozygous deletion identified in K562 by CNVs analysis

Large homozygous deletions (HDs) of recessive cancer genes result in their inactivation and are driver mutations. Several HDs related recessive cancer genes had been found in cancer genomes, including *CDKN2A, RB1, SMAD4, SMARCB1, MAP2K4 and PTEN* (Bignell et al. 2010). An around 6 Mb HD from a gene-rich region on chromosome 9 of K562 was uncovered by CNVs analysis. MCF-7 and HCT116 had normal copy number in the same region indicated that this deletion was not wrongly picked up by mapping error. However, this deletion had no dPET cluster supporting. After close investigation of the boundary of the deletion, we found the repetitive sequence located at the 5' boundary of the deletion, which made it hard for

the dPETs to map. Because of the repetitive sequence, we also could not validate this large deletion by genomic PCR and Sanger sequencing. However, this deletion could be validated by FISH. We used 4 probes, two of them located at the boundary of the deletion (22m4 and 42d9) and the other two located inside the deletion (66p3 and 781g24), and we also used rp11-10619 of chromosome 9 as a positive control. The FISH result clearly showed that the positive control and the two probes located at the boundary of the deletion had signals and the two probes located inside the deletion had no signal (Figure 2.12). There are 37 genes in this 6 Mb HD region, including important tumor suppressor gene *CDKN2A* (Table 2.9). The function of this large HD in K562 need to be further established.

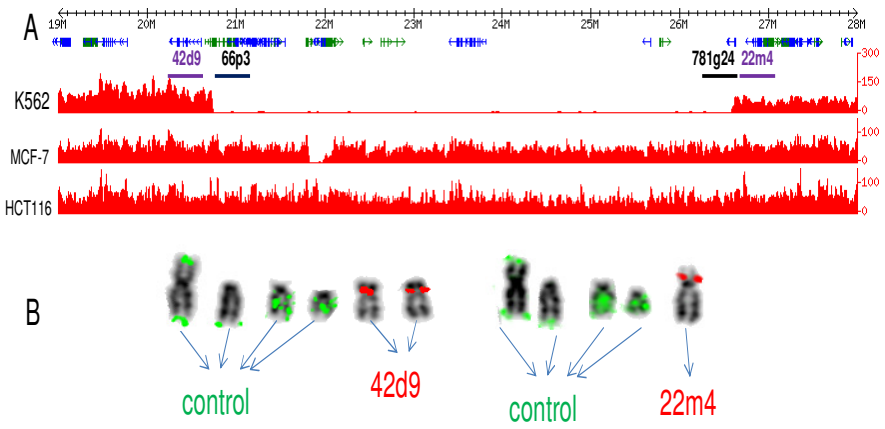

Figure 2.12. A 6 Mb homozygous deletion identified by copy number analysis in K562.

(A) The copy number from 19 Mb to 28 Mb on chromosome 9 in K562, MCF-7 and HCT116. The red track represents the coverage of the cPETs. (B) The FISH validation result of the deletion. Green color was the positive control probe (rp11-106l9: 133087269-133237519, 9q34). Red color was the two probes located outside of the deletion (rp11-42d9: 20560274-20736915, 9p21, rp11-22m4: 26686497-26890847, 9p21). The two probes located inside of the deletion (rp11-66p3: 20784480-20945583 9p21, rp11-781g24: 26409244-26590208 9p21) had no signal.

Table 2.9. Genes located in the HD of chromosome 9 in K562

Chr	Gene Start	Gene End	Gene_span[1)]	Gene_Sym	MCF7 CN[2)]	HCT116 CN	K562 CN
9	23490689	23662385	171696	CR627240	1.5477283	1.814947185	0.144898
9	20993624	21021635	28011	PTPLAD2	1.6929134	2.469914531	0.178242
9	22103664	22111093	7429	BC038540	1.4119948	1.743757167	0.120022
9	21792634	21855969	63335	OK/SW-cl.18	1.0454269	2.096723198	0.174368
9	25666386	25668856	2470	TUSC1	2.0256638	2.41054782	0.240052
9	22436839	22442472	5633	DMRTA1	2.1784669	1.847565768	0.184567
9	21984789	22111093	126304	NR_003529	1.1096129	2.114499375	0.180428
9	21374253	21375396	1143	IFNA2	1.7698008	2.39955456	0.196489
9	21984789	22067889	83100	DQ485454	0.8251085	2.051911687	0.172847
9	21984789	22067889	83100	EU741058	0.8251085	2.051911687	0.172847
9	21314053	21325388	11335	KLHL9	1.5641342	1.7341566	0.159599
9	21314053	21325352	11299	KIAA1354	1.5640891	1.73353197	0.159572
9	21357422	21358075	653	IFNA13	2.7636389	3.505685627	0.3296
9	21470838	21472312	1474	IFNE1	2.0017366	2.324474181	0.225805
9	21130630	21132144	1514	IFNW1	1.3476992	1.739301789	0.160341
9	21444269	21549832	105563	LOC554202	1.7690235	2.399963616	0.217364
9	21294685	21295255	570	IFNA5	1.870899	2.399261484	0.224914
9	25770053	25802963	32910	BC043546	1.6693268	1.914709123	0.194982
9	21992901	21999312	6411	CDKN2B	0.8952875	2.09989774	0.198389
9	21067103	21067943	840	IFNB1	1.6575034	2.177937864	0.18738
9	21196179	21197142	963	IFNA10	4.9191668	6.486508203	0.605519
9	21191233	21229990	38757	IFNA14	4.9191668	6.486508203	0.605519
9	21229200	21229978	778	IFNA14	4.9191668	6.486508203	0.605519
9	21206371	21207310	939	IFNA16	4.9191668	6.486508203	0.605519
9	21217241	21218221	980	IFNA17	4.9191668	6.486508203	0.605519
9	21191467	21192204	737	IFNA7	4.9191668	6.486508203	0.605519
9	21792634	22019593	226959	MTAP	0.6286156	2.173512882	0.217575
9	21430439	21431315	876	IFNA1	1.9597469	2.204164272	0.267893
9	21957750	21984490	26740	CDKN2A	0.3523075	1.537529827	0.147506
9	23680102	23816063	135961	ELAVL2	1.4054098	1.829944442	0.205783
9	21399145	21400184	1039	IFNA8	1.3447016	1.65955186	0.224819
9	22636198	22814212	178014	AX747623	1.6427039	2.334283188	0.277427
9	21267686	21268562	876	V00539	2.1598289	2.779296382	0.338708
9	21155635	21156659	1024	IFNA21	3.5544332	4.186883793	0.502616
9	21176692	21177670	978	IFNA4	3.5544332	4.186883793	0.502616
9	21340316	21340886	570	IFNA6	1.5108888	2.220981059	0.359903
9	20648308	20985954	337646	KIAA1797	1.9466031	2.515487544	0.606575

(1) gene span: gene end minus gene start

(2) CN: copy number

Reconstruction of the *BCR-ABL1* amplicon of K562 by fusion point guided concatenation

The identification of SVs by paired-end sequencing provides a detailed understanding of local genomic structures. However, cancer genomes frequently show complex rearranged amplifications. To reconstruct these complex rearrangements, a collective analysis of the rearrangement points is required. We therefore employed a fusion-point-guided-concatenation algorithm (see Materials and Method) to jointly visualize genomic segments surrounding the translocation (chr9/chr22) which creates the CML causing fusion gene *BCR-ABL1* (Groffen et al. 1984) in K562 (Figure 2.14. D). The analysis showed that i) the disease causing rearrangement point had the highest dPET cluster size (692) supporting the concept of amplified rearrangement points as indicators of driver events (Hillmer et al. 2011); ii) five different chromosomes (1, 3, 9, 13, and 22) were involved in this amplification, and iii) the core region was located in different genomic contexts indicated by alternative paths at the edges of the amplicon (Figure 2.14. A). To place the amplicon picture in a cytogenetic context, we analyzed the three most amplified rearrangement points (largest dPET clusters) by FISH (Figure 2.14. B-C). The FISH analysis confirmed the amplification of the fusion points and showed that the amplicon was distributed over two marker chromosomes. Further, a subpopulation of the *BCR-ABL1* amplicon path connecting chromosome 9 at 133.1 Mb with chromosome 22 at 15.7 Mb (dPET cluster count 218) but not the path connecting chromosome 9 at 133.2 Mb with chromosome 13 at 107.5 Mb was located on chromosome 2q.

We also reconstructed the whole genome rearrangement of MCF-7, HCT116 and K562 by this fusion-point-guided-concatenation algorithm using all the dPET clusters with count ≥ 3 and the detail can be found in the appendix data.

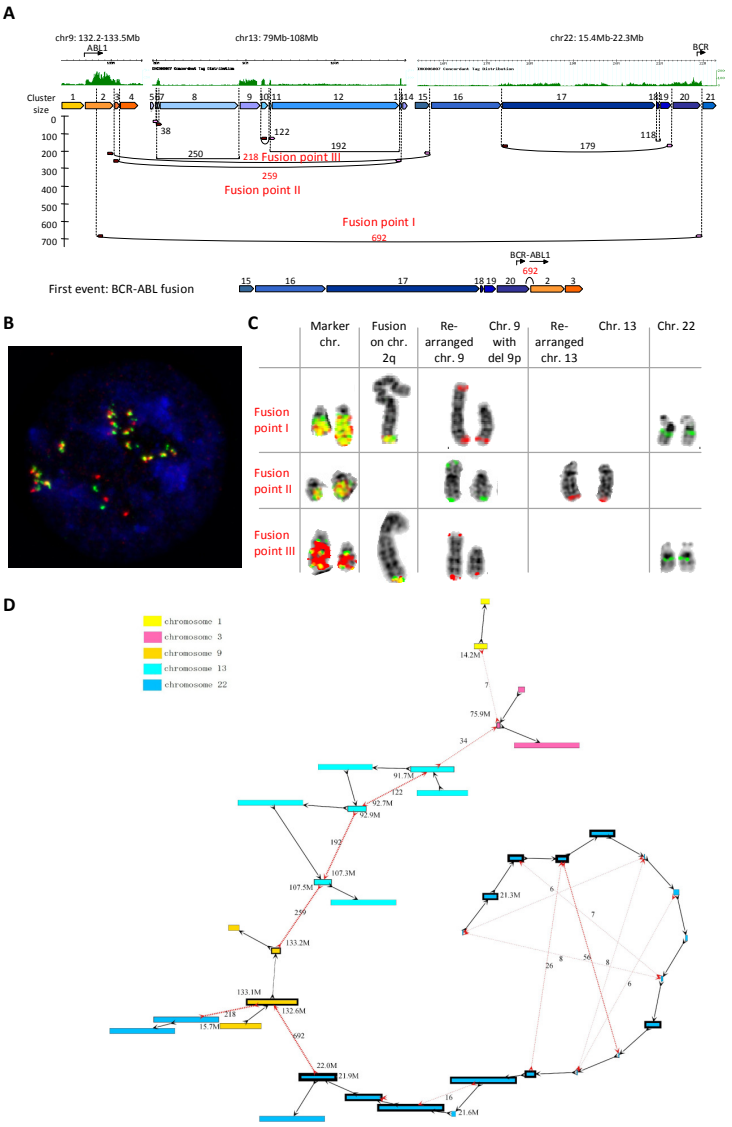

Figure 2.13. Reconstruction of the *BCR-ABL1* amplicon of K562.

(A) Concordant tag distributions representing copy number are shown for amplified genomic regions (top, green track). Genomic segments between predicted breakpoints are indicated by colored arrows and dPET clusters with cluster sizes greater than 35 of predicted somatic rearrangements are represented by horizontal lines flanked by dark red and pink arrows indicating 5' and 3' anchor regions (middle). Small to large dPET clusters are arranged from top to bottom. Cluster sizes are indicated. High dPET cluster size of the CML causing *BCR-ABL1* translocation suggests that the rearrangement occurred early and that it has subsequently been amplified. Fusion points I-III correspond to panel c-d. (B) Fluorescence *in situ* hybridization (FISH) of *BCR-ABL1* rearrangement (fusion point I with cluster size 692). Red RP11-83J21 probe (chr9:132,641,829-132,818,294) and green RP11-61N10 probe (chr22:21,728,292-21,917,898) spanning the fusion point are hybridized on K562 cell line. Yellow spots represent fusion signals and illustrate the amplification of *BCR-ABL1*. (C) FISH analysis of metaphase chromosomes of three high copy fusion points: I) probes as described in b show fusion signals on two marker chromosomes and on chromosome 2q and normal

localization on both rearranged chromosomes 9 and normal chromosome 22; the fusion on chromosome 2 has not been identified by DNA-PET most likely due to low sequence complexity at the break point or complex rearrangements, II) green RP11-106I9 probe (chr9:133,087,269-133,237,519) and red RP11-83J21 probe (chr13:107,393,170-107,577,262) spanning the fusion point II (cluster size 259) show fusion signals on the same marker chromosomes and normal localization on both normal and rearranged chromosomes 9 and 13, III) red RP11-544A12 probe (chr9:132,955,072-133,152,093) and green RP11-104F9 probe (chr22:15,547,686-15,730,740) spanning fusion point III (cluster size 218) show fusion signals on the same marker chromosomes and normal localization on both normal chromosome 22 and rearranged chromosomes 9. (D) Contigs (indicated by boxes) which were covered by PET mapping were concatenated by fusion-point-guided-concatenation method. The length of a contig is represented by the length of the box. Because of the size difference between chromosomes 1, 3, 9, 13, and 22, the length of chromosome 22 is represented by the length of contig/10,000 while the lengths of chromosomes 1, 3, 9, and 13 are represented by the length of contig/100,000. Any value less than 0.1 is rounded to 0.1; any value larger than 6 is rounded to 6. The thickness of borders of each contig represents the coverage (copy number). Red dashed edges represent dPET edges, while black bold edges represent cPET edges. The thickness of dPET edges represents the size of the corresponding dPET cluster. cPET edges have uniform thickness. Arrow heads pointing towards a contig indicate connections with the lower coordinates, arrow heads pointing away from a contig indicate connections with the higher coordinates.

DISCUSSION

The PET sequencing has become a key technique to assess genome rearrangements and SVs in normal and cancer genomes (Clark et al. 2010; Fujimoto et al. 2010; Hillmer et al. 2011; Lee et al. 2010; McKernan et al. 2009; Pleasance et al. 2010b; Stephens et al. 2009). However, some characteristics are still not well understood regarding study design and the balance of cost versus benefit of different sequencing strategies. One such factor is the choice of the most suitable sequencing library insert size. Using a quantitative study, Bashir *et al*. concluded that larger clones could maximize the clonal coverage and detect as many rearrangement breakpoints as possible while reducing the sequence effort, whereas smaller clones could provide better localization (Bashir et al. 2008). This conclusion was confirmed by Bentley and McKernan, who observed that when using different insert sizes, most of their predictions were unique to one data set, and the probability of detecting a breakpoint

with a combined library was higher than using only one type of library (Bentley et al. 2008; McKernan et al. 2009). However, 200 bp to 3 kb insert sizes are not large enough since only lengths of > 7 kb are expected to span common transposon insertions such as L1s (the canonical L1 element is 6 kb long) (Cordaux et al. 2009) and thereby can identify insertion events in a single read. Using three cancer cell lines, MCF-7, HCT116, and K562 as test genomes, we sequenced different insert size libraries (1 kb, 10 kb, and 20 kb) to identify SVs. The comparison of different insert size libraries demonstrated that the PET sequencing strategy with large insert sizes (10 kb) is an attractive whole-genome sequencing approach to identify SVs in human genomes. With the same sequencing effort, the 10 kb libraries could identify more and larger SVs, whereas the 1 kb libraries were advantageous in identifying deletions with span < 5 kb. In addition, 10 kb libraries had a comparable resolution in predicting breakpoint locations to a distance that can be amplified by PCR. The 20 kb insert size library had a slight advantage in discovering inversions and unpaired inversions but displayed a lower sensitivity in identifying small SVs of various categories compared to 10 kb insert size library. The construction of libraries with 20 kb inserts requires more genomic DNA starting material compared to 10 kb insert libraries. The detailed characterizations of SVs by large insert size libraries showed many new sub-types of insertions, which could help in understanding the genesis and effect of insertions in human normal and cancer genomes.

This study is complementary to those that have investigated the effect of read-lengths and library-size on the ability to do *de novo* assembly of the data (Chaisson et al. 2009; Nagarajan et al. 2009; Wetzel et al. 2011). In a recent work (Wetzel et al. 2011) the authors suggest that multiple library sizes are needed to optimally resolve various classes of repeats. While larger library sizes allow the spanning of more

repeat classes the associated complexity of assembly analysis also increases. These considerations make the choice of library-size for *de novo* assembly less clear-cut when compared to reference-guided SV analysis.

With the rapid development of next generation sequencing technologies, whole genome sequencing has become an invaluable tool for obtaining a complete understanding of human genomic variation. In the future, personal genomic information will gain importance to tailor an individual's medical care. Our study provides valuable information on the characteristics of PET sequencing libraries and such information will help to select appropriate and most effective insert sizes for various kinds of sequencing projects.

Chapter Three: Long Span PET Mapping Reveals Characteristic Patterns of Structural Variations in Epithelial Cancer Genomes

Introduction

Many important cancer genes have been identified at translocation breaks in leukemias, lymphomas and sarcomas, by contrast, the possibility that fusion genes might be present in the common human epithelial cancers, such as colorectal and breast carcinoma, was largely ignored (Edwards. 2010). We have a relatively good understanding of the genes which can be point-mutated, amplified or deleted in these cancers, however, the large number and the complexity of genome rearrangements has made it difficult to identify all genes at chromosome breakpoints. It has been widely assumed that gene fusions do not contribute significantly to carcinomas, however, the identification of *TMPRSS2-ERG* in prostate cancer (Tomlins et al. 2005) and *EML4-ALK* in lung cancer (Soda et al. 2007) confirmed that fusion genes are also present in solid tumors, but that their detection has been hampered for technical reasons.

The limited knowledge of chromosome rearrangement in the common cancer is due to the lack technologies which allow a comprehensive analysis. Traditional cytogenetic approaches require metaphase chromosome preparations and therefore cell lines which are not easy to establish from a large number of primary tumors. In addition, the complexity of carcinoma karyotypes renders the identification of rearrangements by cytogenetic banding challenging and prone to errors (Adeyinka et al. 2000). Until the last year or two, it has not been technically feasible to systematically find genomic rearrangements. The genomic rearrangements identified in carcinomas were found by indirect methods that can be applied only for particular fusions. However, two complementary approaches led us to a new era. One approach derived from molecular cytogenetics, i.e. FISH and CGH; while the other uses new sequencing technologies.

The introduction of a new generation of DNA sequencers has provided a different way to systematically discover genome rearrangements. There are two approaches to discover genome rearrangements by DNA sequencing. In the first approach, fragmented genomic DNA or cDNA can simply be sequenced at random and can be compared to the reference genome and transcriptome. This has worked very well when applied to cDNA (Guffanti et al. 2009) but it is too expensive when applied to the whole human genome. The second approach is "paired end read", where only the two ends of long genomic DNA fragments are sequenced. The paired end sequencing approach is very well suited for the discovery of genomic rearrangements. A particular strength is that it can be applied to DNA of both, tumor and cell lines. Recently 15 breast primary tumor and 9 immortal breast cancer cell lines had been sequenced by paired-end sequencing in order to understand the patterns of somatic rearrangement in breast cancer genomes (Stephens et al. 2009). A total of 2,166 confirmed somatic rearrangements were identified among the 24 cancers. However, there is substantial variation the frequency of somatic rearrangements across cancer samples. Overall, breast cancer cell lines showed more rearrangements (median 101, range 58-254) than primary tumors (median 38, range 1-231). This difference may be due to the acquisition of additional rearrangements during *in vitro* cultures or the contamination of normal tissue which renders the detection of somatic events more difficult or the relative propensity of some subclasses of breast cancer to become established in culture.

Many novel in-frame fusion genes or internally rearranged genes were identified and most of them were expressed. However, none of them was recurrent. It is more likely that most are passenger events and much larger series will be required to investigate comprehensively the possibility of recurrent cancer-causing rearrangements in breast cancer. This study gives insight into the complexity of rearrangement patterns in solid tumor genomes and shows that most

rearrangements in breast cancer are intrachromosomal. Breast cancers are highly heterogeneous and the prevalence of tandem duplications can be used to subclassify breast cancer. Breast cancers with many tandem duplications are usually oestrogen- and progesterone-receptor negative and classified by expression profile as basal-like. In contrast, cancers with few rearrangements or with rearrangements in amplicons are usually oestrogen-receptor positive and are classified as luminal A and luminal B types, respectively (Stephens et al. 2009). Although this study was restricted to breast cancer, many of the findings will also be relevant for other common cancers, and they are consistent with a preceding pilot study of two lung cancer cell lines (Campbell et al. 2008).

Recently, five studies described the application of genome analysis techniques to a range of breast cancer. Curtis and colleagues analyzed copy number, sequence changes known as single nucleotide polymorphisms, and gene-transcription rates in approximately 2,000 breast cancers which included all known breast cancer types (Curtis et al. 2012). Around 40% of genes' expression were associated with inherited variants (copy number variants and single nucleotide polymorphisms) and acquired somatic copy number aberrations (CNAs). Three putative cancer genes were identified, including deletions in *PPP2R2A*, *MTAP* and *MAP2K4*.

In order to comprehensively understand the driver mutations and mutational processes operative in breast cancer, the genomes of 100 tumors for somatic copy number changes and mutations in the coding exons of protein-coding genes had been sequenced (Stephens et al. 2012). The results showed that at least 40 cancer genes were implicated in the development of the 100 breast cancers, including point mutations and/or copy number changes. The maximum number of mutated cancer genes in an individual cancer was 6, but 28 tumors only showed a single driver. In some cases, the presence of multiple drivers was associated with subclonal evolution of the

cancer. Several new cancer genes were identified, including *AKT2*, *ARID1B*, *CASP8*, *CDKN1B*, *MAP3K1*, *MAP3K13*, *NCOR1*, *SMARCD1* and *TBX3*. Among the 100 tumors, 73 different combinations of mutated cancer genes had been found and this strongly highlighted the substantial genetic diversity underlying this common disease.

Primary triple-negative breast cancer (TNBC) is a tumor type defined by lack of oestrogen receptor, progesterone receptor and ERBB2 gene amplification, which represent approximately 16% of all breast cancers. Shah *et al.* assessed mutations, cop number and gene expression in 104 TNBC cases and found that the frequencies of copy number abnormalities and mutations vary markedly between and within the tumors, which indicates that mutations can arise at multiple stages of tumor progression (Shah et al. 2012). Three genes including TP53, PIK3CA and PTEN, are involved in the early stages of breast-cancer development. Interestingly, only one-third of the low-prevalence mutated genes identified by this study were transcribed into RNA, which suggests that they may be chance mutations unrelated to the cancer and/or the mutations involved genes with tumor-suppressive activity.

Ellis and colleagues studied pretreatment tumor biopsies accrued from patients in two studies of neoadjuvant aromatase inhibitor therapy to correlate the variable clinical features of oestrogen-receptor-positive breast cancers with somatic alterations (Ellis et al. 2012).Eighteen significantly mutated genes were identified, among of them, five genes (*RUNX1*, *CBFB*, *MYH9*, *MLL3* and *SF3B1*) were reported to be related with haematopoietic disorders. The researchers also showed that compared to the tumor cells with lower Ki67 protein expression, the tumor cells with a higher expression level were associated with resistance to aromatase inhibitors and contained more somatic mutations and genome structural changes. This finding implicated

genetic changes which lead to deregulation of DNA replication and repair processes in this drug resistance.

In the last study, Banerji and colleagues examined whole-exome sequences of DNA from 103 human breast cancers of diverse subtypes and whole-genome sequences of 22 breast cancer/normal pairs (Banerji et al. 2012). Besides confirming recurrent somatic mutations in *PIK3CA*, *TP53*, *AKT1*, *GATA3* and *MAP3K1*, recurrent mutations in the *CBFB* transcription factor gene and deletions of its partner *RUNX1* were also discovered. Moreover, a recurrent *MAGI3-AKT3* fusion was identified to be enriched in TNBC. The MAGI3-AKT3 fusion led to constitutive activation of AKT kinase, which is abolished by treatment with an ATP competitive AKT small-molecule inhibitor.

Gastric cancer is also a heterogeneous disease with multiple environmental etiologies and alternative pathways of carcinogenesis. Besides mutations in *TP53*, changes in other genes or pathways only account for small subsets of the disease. In order to identify previously unreported mutated genes and pathway alterations, 22 gastric cancer samples had been studied by exome sequencing (Wang et al. 2011) and the result showed that genes involved in chromatin modification to be commonly mutated. Furthermore, the mutation spectrum for *ARID1A*, which encodes a member of the SWI-SNF chromatin remodeling family, differed between molecular subtypes of gastric cancer and mutation prevalence was negatively associated with mutations in TP53.

To explore the spectrum of somatic mutations in gastric cancer, the exomes of 15 gastric adenocarcinomas and their matched normal DNAs were sequenced (Zang et al. 2012). The frequently mutated genes included *TP53*, *PIK3CA* and *ARID1A*. Cell adhesion was the most

enriched biological pathway among the frequently mutated genes. Around half of the gastric cancers had mutations in chromatin remodeling genes (*ARID1A, MLL3 and MLL*) and 8% of primary tumors had *ARID1A* mutations. A cadherin family gene, *FAT4*, was confirmed to be mutated in 4% of gastric tumors. The functional assays showed that both *FAT4* and *ARID1A* to exert tumor-suppressor activity. Somatic inactivation of *FAT4* and *ARID1A* may be the key tumorigenic events in a subset of gastric cancers.

All these studies are remarkable testament to the power of genomic technologies to define the landscape of a complex disease state, and give us the most thorough view yet of the molecular underpinnings of breast and gastric cancers. In this chapter, we applied the 10 kb long span paired-end-tag sequencing and mapping strategy to two epithelia cancers, i.e. breast and gastric cancer. We comprehensively mapped genome SVs of eight breast cancer samples including five primary breast cancer tumors and three well established cell lines (MCF-7, T47D, and SKBR3), five gastric cancer samples including four primary gastric cancer tumors and one cell line (TMK1). These are contrasted to genomes of a colon cancer cell line (HCT116), a chronic myelogenous leukemia (CML) cell line (K562), and 2 normal individuals (an African and a European). Cross comparison of the cancer and normal genomic maps enabled us to distinguish possible somatic rearrangements from germ line events, and revealed characteristic patterns of SVs that are prominent in breast and gastric cancer genomes. Using the connectivity and quantitative nature of the DNA-PET data, we delineated the genealogy of rearrangement events involved in amplified regions in individual cancer genomes, and elucidated potential underlying mechanisms involved in cancer genome instability and aneuploidy.

Results

Genomic DNA-PET sequencing and mapping

The genomic DNAs of 17 human genomes were sheared randomly and gel purified in 10 kb size range (Figure 3.1, breast tumor 13 DNA was purified in 5 kb range due to limited DNA quality). The DNA fragments of each genome were processed for PET construction and paired-end sequencing analysis. In total we generated >25.9 Giga bases of DNA sequence derived from >476 million non-redundant PET sequences from these 17 genomes and achieved, on average, 81-fold physical (fragment) coverage of each genome (Table 3.1). Five libraries, including the European normal genome, MCF-7, two gastric tumors and TMK1, have achieved more than 100-fold physical coverage due to either larger fragment size or more sequence reads.

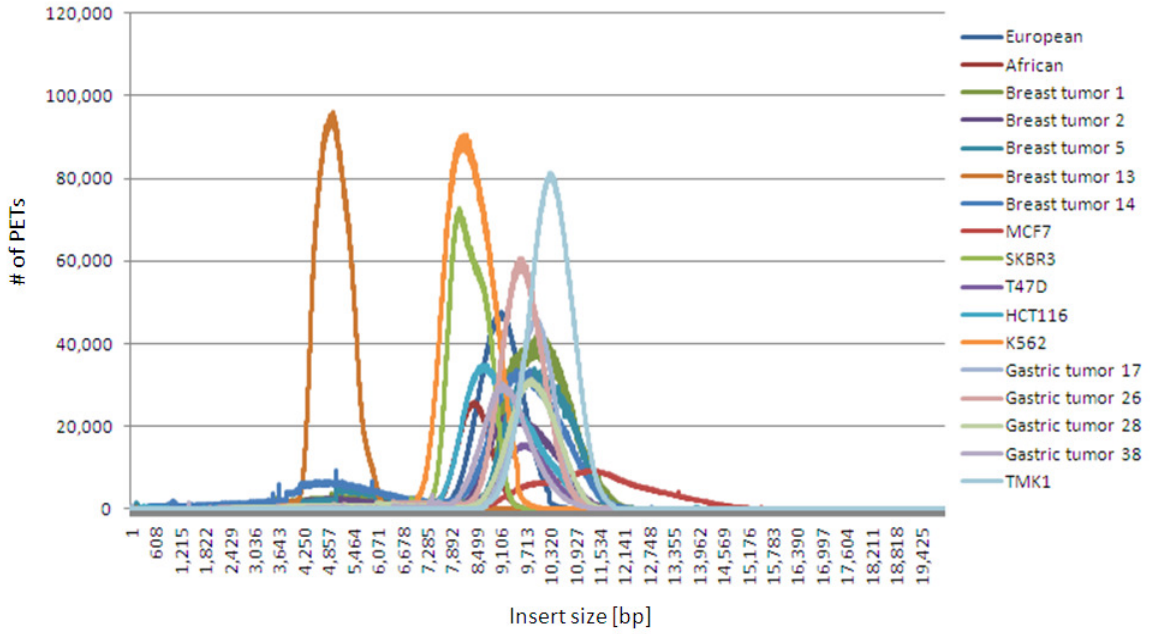

Figure 3.1. DNA-PET libraries insertion size distribution.

DNA insertion size peaks range from 5 kb (breast tumor 13) to 11.5 kb (MCF-7). The majority of libraries show inserts with a size of 10 kb. Breast tumor 14 showed PCR caused spikes which have been excluded for this presentation.

Classifying cPET and dPET

With the continuous optimization of library construction protocol, the libraries constructed at the late stage of this project had much sharper library span compared to the libraries constructed at the early stage (Figure 3.1). Therefore, in some cases the gradient used to determine the minimum span point of particular library (see Materials and Methods for Chapter 2) did not reach 0 or reached 0 much further from the expected point. Therefore we have defined the cutoff gradient as 0.01 x Maximum gradient. The first point at which this gradient occurred was considered as the minimum span of the libraries. Similarly the maximum span was determined as the first point to the right of the minimum gradient point where the gradient reached 0.01 x Minimum gradient (Figure 3.2). This definition gave very similar cutoffs compared to the SOLiD pipeline Corona Lite version 4.0.2 cutoff for the rescue window which was introduced at the later stage of the project. The Corona Lite version 4.0.2 cutoff definition was used for libraries IHT009 and DHG003. The vast majority of PET sequences (89%) were cPETs and the density of the cPETs in any region of the genome was used to reveal chromosomal copy number variations.

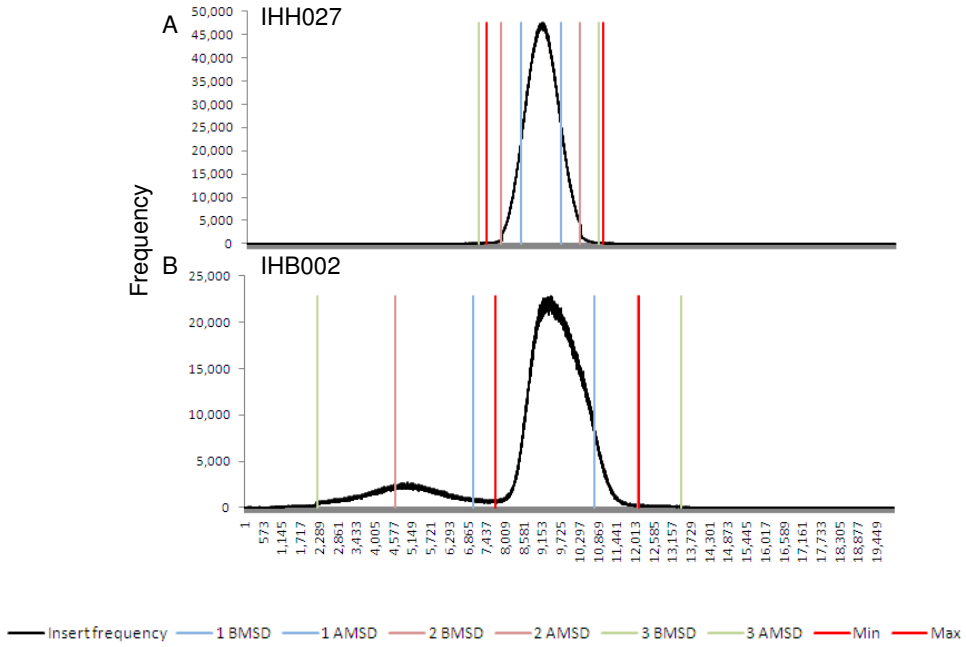

Figure 3.2 Gradient based span cutoff compared to standard deviation based cutoff.

Frequencies of library insert sizes for two DNA-PET libraries are shown in black (IHH027, European; IHB002, Breast tumor 2). Below and above median standard deviations (BMSD and AMSD, respectively) are shown in blue, brown, and green for 1, 2, and 3 standard deviations, respectively. Gradient based definition of minimum and maximum cPET are shown in red. Note that for IHB002, 2 AMSD (12,112) and the maximum cPET cutoff (12,163) are superimposed.

Table 3.1 Statistics of massively parallel PET sequencing of each genome.

Sample	Tags	Mappable Tags	PET	PET (NR)[1]	cPET[2] span range [bp]	cPET (NR)[1]	Coverage[3]	dPET[4] (NR)[1]	dPET[4] cluster[5]
European	366,708,248	200,414,308	64,350,297	43,646,070	7,400-10,977	40,682,877	130.1	2,963,193	666
African	251,094,242	122,504,474	34,441,940	27,013,890	7,132-9,854	24,075,215	71.5	2,938,675	356
Breast tumor 1	353,330,838	217,285,714	86,355,790	5,454,539	8,323-12,436	4,705,578	16.8	748,961	313
Breast tumor 2	217,007,124	134,511,878	50,962,000	10,951,889	7,742-12,163	10,214,722	34.9	737,167	426
Breast tumor 5	496,450,386	280,659,839	78,184,988	10,276,850	8,109-12,120	6,710,288	23.8	3,566,562	242
Breast tumor 13	580,449,070	335,245,735	108,789,285	38,594,882	4,012-6,325	34,746,195	61.4	3,848,687	957
Breast tumor 14	727,464,383	355,410,533	95,183,702	22,615,058	7,592-12,520	16,603,012	57.2	6,012,046	434
MCF-7	362,405,602	150,722,313	35,438,575	28,518,121	8,099-16,217	25,208,550	101.1	3,309,571	1,047
SKBR3	357,021,381	228,828,267	83,688,064	25,624,846	7,229-9,639	23,253,608	68.4	2,371,238	1,145
T47D	177,908,044	90,297,366	26,564,983	14,718,535	7,816-11,617	12,997,188	44.5	1,721,347	376
Gastric tumor 17	622,317,436	234,345,611	53,012,028	42,039,629	8,630-11,310	40,195,339	140.5	1,844,290	1,126
Gastric tumor 26	407,475,701	250,309,081	84,647,235	14,603,105	8,108-11,653	11,798,863	40.1	2,804,242	586
Gastric tumor 28	270,976,929	157,958,274	51,743,377	32,832,641	8,152-11,786	30,557,538	106.3	2,275,103	1,238
Gastric tumor 38	351,714,141	194,346,844	48,868,556	28,728,468	7,561-11,503	24,491,622	79.8	4,236,846	550
TMK1	501,758,330	306,105,871	110,164,772	67,903,728	8,760-11,920	64,162,963	233.3	3,740,765	1,253
HCT116	492,180,222	292,721,687	72,893,710	28,262,465	7,200-11,780	24,599,666	77.0	3,662,799	883
K562	376,077,182	223,850,946	133,675,193	34,759,699	6,846-10,248	31,240,240	91.9	3,519,459	939

[1)] non redundant
[2)] concordant PET
[3)] physical coverage
[4)] discordant PET
[5)] clusters of size ≥ 3

Purifying dPETs clusters

Among the dPETs (11% of PET sequences), the most prominent dPET categories were small span PETs (median of 17 libraries = 4.07% of total PETs) and different chromosome PETs (median of 17 libraries = 4.53% of total PETs; Figure. 3.3). Some libraries (i. e. constructed at the beginning of this project with larger standard deviation had a large fraction of small span PETs (Table 3.2) with the extreme of 21.43% of all PETs having a small span for breast tumor 14. All small span PETs were excluded from further SVs analysis since they interfered with the clustering procedure.

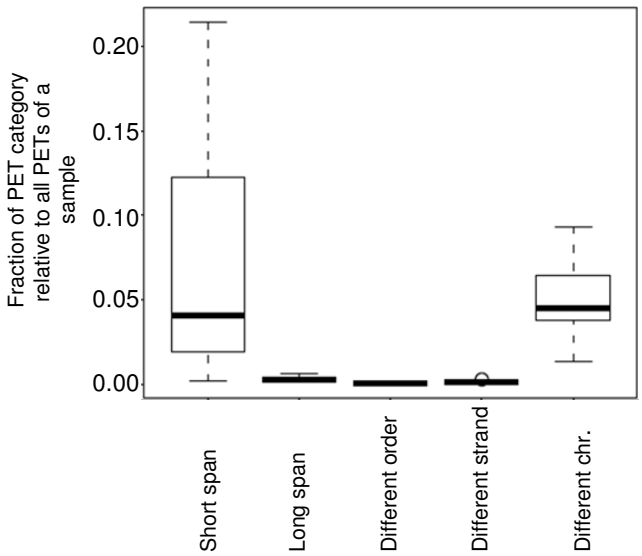

Figure 3.3. Frequencies of non-cPET categories of 17 DNA-PET libraries.

Non-cPETs of each library were categorized in the five shown categories as described in Materials and Methods of Chapter 2 'PET classification'. Fraction of PET counts of each category compared to total PET counts per sample is presented. Fractions of 17 libraries are summarized by box plots. Thick horizontal lines indicate medians, boxes represent values of libraries from the 25th to 75th percentile, whiskers indicate minimum and maximum.

Table 3.2: Median and standard deviations of DNA-PET library inserts

Library	Sample	Median [bp]	Below-Median Standard Deviation [bp]	Above-Median Standard Deviation [bp]
IHH027	European	9,103	636	577
IHH022	African	8,294	2,697	1,281
IHB001	Breast tumor 1	9,909	2,271	1,183
IHB002	Breast tumor 2	9,449	2,403	1,332
IHB005	Breast tumor 5	9,785	2,931	1,590
IHB013	Breast tumor 13	4,954	974	548
IHB014	Breast tumor 14	9,338	3,318	2,046
IHM005	MCF7	10,198	2,515	1,314
IHM006	MCF7	11,972	2,779	1,624
IHS012	SKBR3	8,279	815	524
IHT008	T47D	9,599	2,217	1,038
DHG003	Gastric tumor 17	10,011	848	505
IHG009	Gastric tumor 26	9,561	1,997	899
IHG005	Gastric tumor 28	9,785	2,396	1,111
IHG006	Gastric tumor 38	9,188	1,726	867
IHT009	TMK1	10,347	653	562
IHH003	HCT116	9,711	3,383	1,710
IHH020	HCT116	9,282	1,140	685
IHH026	HCT116	8,513	562	526
IHK006	K562	8,305	643	594
IHK007	K562	8,304	655	597

During construction of a DNA-PET library, chimeric ligation products occur and it is more likely that two randomly ligated genomic DNA fragments are of different chromosomes than from the same chromosome. Similarly, dPETs creating mapping artifacts due to sequence similarity between two different genomic regions are more likely to occur inter-chromosomally. In the 17 DNA-PET libraries we observed >56 million different chromosome PETs (including redundant reads) representing >90% of all dPETs after excluding the short span category. Only 23,900 different chromosome PETs formed 999 inter-chromosomal clusters, representing 8% of all dPET clusters. We also conducted a simulation study where we generated 100 random datasets of chimeric PETs (14.6 million of them = 11% of 133 million, the number of PETs in the largest library in this study [K562]) and clustered them. While on average, each dataset had nearly 130 size 2 clusters, only one of the datasets had a single, count 3 cluster, which strongly confirming that such artifacts are unlikely in our results when using cluster count 3 as cut off.

For libraries IHB002, IHB005, IHB014, IHH022, and IHT008, the span window for the rescue mapping procedure (detail see Material and Methods for Chapter 2) was larger than the window used to define the cPET cutoff. These rescued PETs caused large numbers of artifactual dPET clusters of the category: same chromosome, same strand, correct ordering and larger span distance. These artifacts were removed by excluding from further analysis all clusters of the category same chromosome, same strand, correct ordering with a distance smaller than the rescue window between the start of the 5' anchor and the start of the 3' anchor (corresponding to StartL and StartR in appendix data, Table 13) or the end of the 5' anchor and the end of the 3' anchor (corresponding to EndL and EndR in appendix data, Table 13). dPET clusters with anchor regions <1 kb for 10 kb libraries and <500 bp for the 5 kb library of breast tumor 13 and clusters within 1 Mb of centromeres were excluded from further analysis.

Detection of rearrangement points and structural variations (SVs)

After dPETs clusters purification and using cluster count 3 as cut off (Figure 3.4), the numbers of rearrangement points (the junction of two genomic breakpoints) identified by dPET clusters in the 17 genomes ranged from 242 in breast tumor 5, to 1,255 in the gastric cancer cell line TMK1 (Table 3.1). By plotting the increasing curve of dPET clusters against the sequencing depth (non-redundant PETs), it is estimated that with 27 million or more non-redundant PETs, we would be able to identify approximately 80% of SVs that could be discovered by this technology (Figure. 3.5). We calculated that using standard short tag sequencing strategies with 500 bp fragments; we would require 540 million non-redundant PET reads to match this threshold. Most of the 17 genome datasets were either above or close to this mark, except that of 3 breast tumors (BT1, BT2, BT5) which had only 10 or 5 million non-redundant PET sequences for approximately only 40-50% of SVs. In our later analyses, we noted that these three tumors showed under representation of SVs most likely due to the lack of comparable coverage. In order to discriminate the simple and complex SVs in each genome, supercluster analysis was done in each 17 genomes and all the SVs which had supercluster count ≥4 were identified as complex SVs (Figure 3.6).

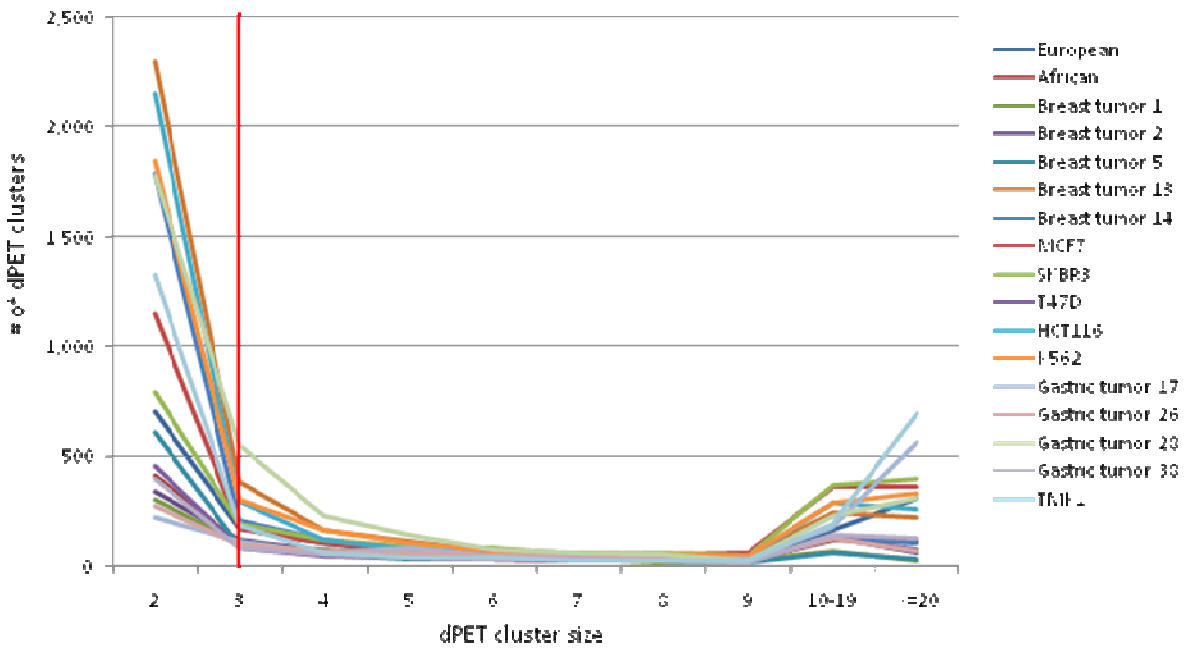

Figure 3.4. dPET cluster count distribution of 17 DNA-PET libraries.

The number of observed dPET clusters (y-axis) is shown for the individual cluster counts (x-axis). Red vertical line represents the cutoff for dPET clusters regarded as reliable breakpoint pairs (count three and higher).

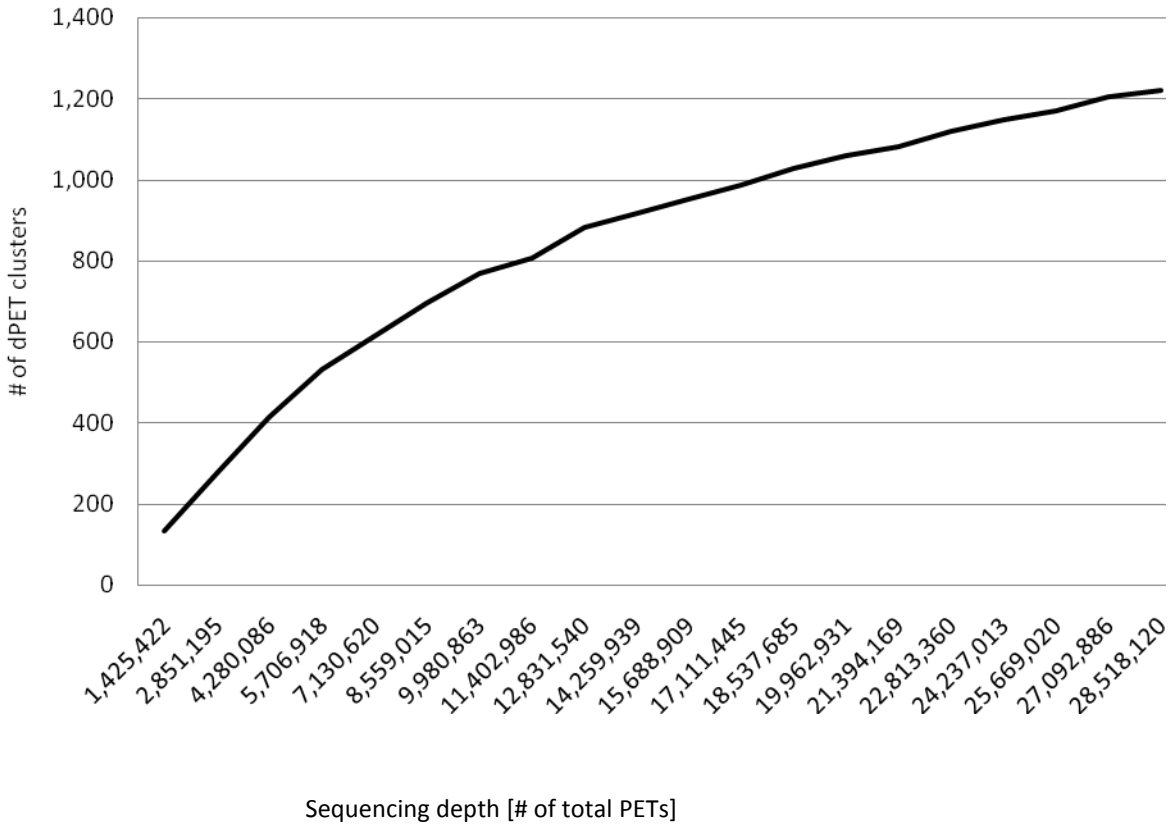

Figure 3.5. Saturation curve for breakpoint discovery.

The total of all non-redundant PETs (x-axis) of MCF-7 was reduced in increments of 5% and plotted against the resulting number of dPET clusters of size three and higher (y-axis).

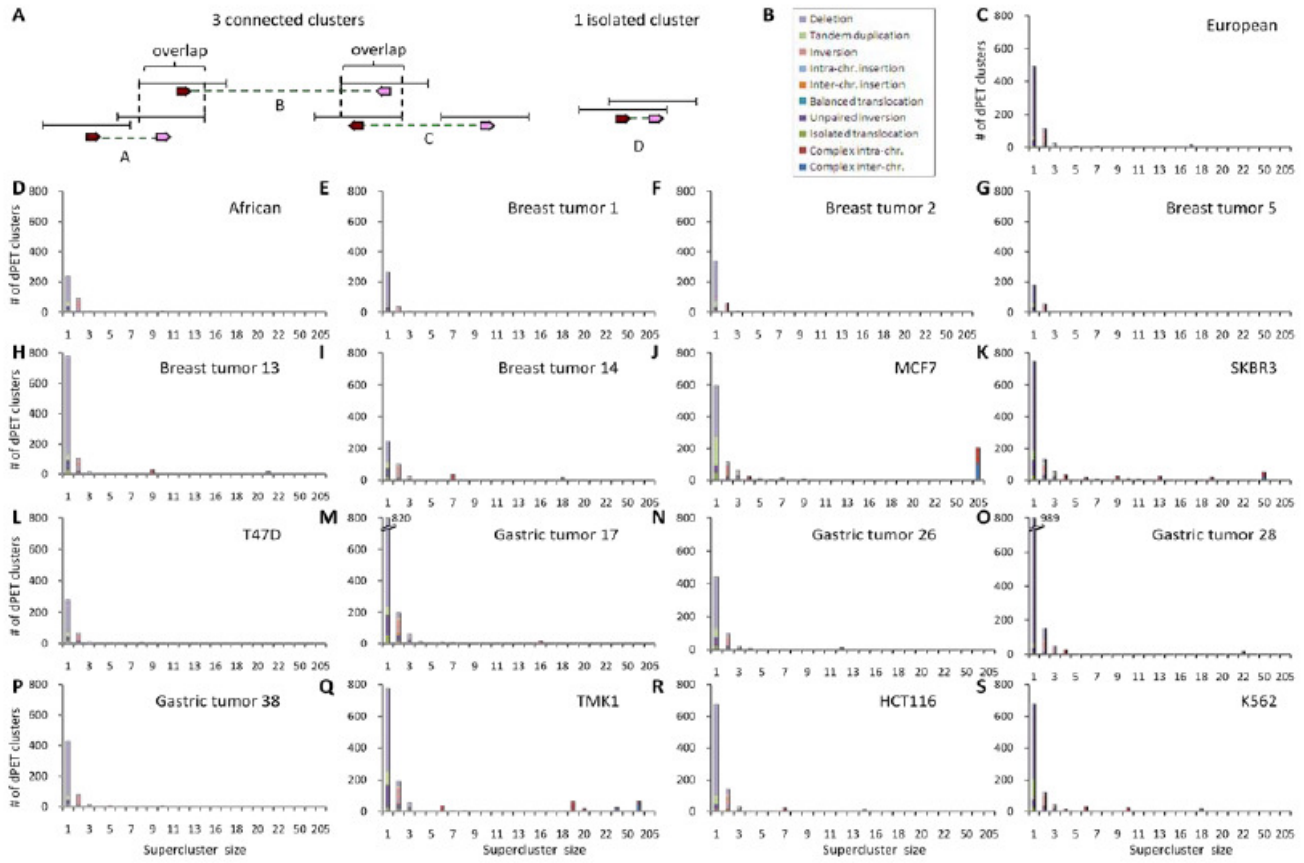

Figure 3.6. Connectivity of breakpoints in 17 DNA-PET libraries.

(A) Interconnection network establishment. (B) Color code for c-s. (C-S) Distribution of degrees of connectivity for the indicated genomes (top right). Numbers of clusters (y-axis) for each supercluster size (number of interconnected clusters, x-axis) is shown.

In total, we predicted 12,537 SVs in these 17 genomes (Table 3.3, Details in Appendix Table 4). Majority of them (*n*=11,394, 91%) were simplex SVs and only 9% (*n*=1,143) were complex SVs. Three genomes including MCF7, SKBR3 and TMK1 had most abundant complex SVs more than two hundred (Table 3.3). The genomes which had the complex SVs less than ten were Breast tumor 1, 2 and 5 and this might due to the lack of comparable coverage, because the three tumor samples had the lowest coverage among the 17 genomes. Deletion is the most predominant SV type (*n*=7,376; 59%), ranging from 131 deletions found in breast tumor 5 to 1,027 in gastric tumor 28. Each the other SV types accounted for less than 10%. Among them, unpaired inversions (*n*=1,222, 9.7%), complex rearrangements (n=1,143, 9.1%), tandem duplications (n=1,033, 8.2%), and inversions (n=908, 7.2%) were considered at the same frequency level, whereas insertions (intra-chromosomal n=264, 2.1%; inter-chromosomal n=105, 0.8%) were in lower frequency. We also detected a significant number of isolated inter-chromosomal translocation events (n=482, 3.8%), but very few balanced translocations, one (2 rearrangement points) in T47D and one in TMK1.

The span of the predicted SVs ranged from 402 bp to >216 Mb. One hundred and sixty-four deletions were predicted to have a size below 2 kb with the smallest of 402 bp. Since deletions are called based on the stretched mapping and this is dependent on the distribution of the insert size of a library, predicted deletions <2 kb are increased for false positives. One hundred and fifty-eight deletions were predicted to have a span >1 Mb with the largest of >216 Mb. It is possible that other events inside large SVs change the overall picture although the connectivity network established by the superclustering reduces the rate of misclassification of SVs by describing regions with many breakpoints as complex. The smallest predicted tandem duplication had the size of 1,132 bp, the largest was >127 Mb. The smallest predicted inversion

Table 3.3: Isolated and complex SVs in 17 DNA-PET libraries

		Deletion	Tandem duplication	Unpaired inversion	Inversion	Intra-chr insertion	Inter-chr insertion	Isolate translocation	Balanced translocation	Complex intra-chr	Complex inter-chr	Physical coverage
Normal	European	460	28	43	58	20	6	18	0	18	15	130
	African	189	25	38	64	16	0	14	0	2	8	72
Breast tumor	BT1	226	13	31	26	6	0	11	0	0	0	17
	BT2	280	40	32	38	10	6	13	2	2	5	35
	BT5	131	22	24	38	8	2	12	0	0	5	24
	BT13	682	39	85	38	22	6	34	0	33	18	61
	BT14	159	46	70	56	16	7*	22	0	40	18	57
Breast cancer cell line	MCF-7	352	203	83	60	14*	8	59	0	146	122	101
	SKBR3	606	78	135	56	14	6	45	0	158	47	68
	T47D	223	32	37	40	6	0	24	2	6	6	45
Gastric tumor	GT17	640	71	180	80	19*	14	77	0	32	13	141
	GT26	341	62	68	48	15*	4	28	0	6	14	40
	GT28	1027	29	38	62	16	6	14	0	31	15	106
	GT38	375	28	40	54	15*	4	14	0	12	8	80
Gastric cancer cell line	TMK1	571	103	183	76	27*	11*	48	2	150	82	233
Other cancer cell lines	HCT116	614	74	49	58	24	12	13	0	26	13	77
	K562	500	140	86	56	16	13*	36	0	73	19	92

* one of the two clusters belong to another SV category

was 996 bp, the largest >115 Mb. Insertions ranged from 1,165 bp to >77 Mb and unpaired inversions had a distance between the predicted breakpoints of 555 bp to >216 Mb. Overall, there is technically no upper size limitation to detect SVs by 10 kb insert DNA-PET. We did not undertake extensive efforts to evaluate the lower boundary to detect small deletions but the discovery rate is lower and false discovery rate higher for the deletions <2 kb. The lower size boundary for other SV categories is limited by our minimum anchor size requirement of 1 kb for 10 kb DNA-PET libraries and 500 bp for the 5 kb library (Breast tumor 13).

Interestingly, the median cluster count of the 17 genomes were higher for deletions (median = 11) and inversions (median = 13) compared to the cluster count of the other SVs (medians ranging from five to eight; Figure 3.7). Unpaired inversions showed the lowest median cluster size of five. The larger median cluster size of deletions correlated with a higher PCR validation rate (see below) and was mainly due to deletions smaller than 10 kb. Inversions had a high rate of repeated observations across the 17 genomes which increases the confidence and also indicates that many inversions are of high frequency. High frequency events are more likely to be in a homozygous state which will result in a larger cluster size than heterozygous events. Rearrangement points of somatic events, if not amplified, are expected to be at a lower copy number per tissue sample than a homozygous germline SV due to heterogeneity across cells within a tumor. We investigated the median cluster size of the tumor samples vs. the normal samples and cell lines and indeed found a trend of lower cluster sizes for tumor SVs compared to the (more) homogenous normal and cell line samples (Figure 3.7).

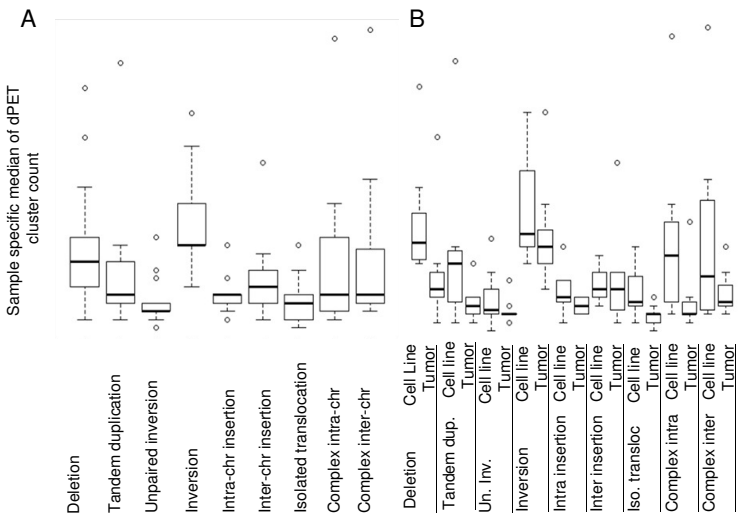

Figure 3.7. Median of SV specific dPET cluster counts for 17 DNA-PET libraries.

SV categories are shown on the x-axis, and the median of the dPET cluster count for each genome is shown on the y-axis. Thick horizontal lines indicate median, boxes represent values of libraries from the 25th to 75th percentile, whiskers indicate minimum and maximum, and circles indicate outliers. (A) Box plots of 17 genomes; (B) box plots of 17 genomes divided into cell lines and normals (n=9) and tumors (n=9).

Collectively, the concordant and discordant DNA-PET mapping data constitute the comprehensive SV map for each cancer and normal genome (Figure 3.8), displaying the precise genome architecture and quantitative measurements of copy number variations. For example, the SV map of K562 showed the accurate position of the known *BCR-ABL1* translocation between chromosome 9 and 22 (Daley et al. 1990; Groffen et al. 1984; Heisterkamp et al. 1990), and the previously reported *BCAS3-BCAS4* fusion between chromosome 17 and 20 in MCF-7 (Ruan et al. 2007). The two breast cancer cell lines SKBR3 and MCF-7 showed the most extreme amplification hotspots for inter-chromosomal rearrangements (e.g. between chromosomes 8, 14 and 17 in SKBR3 and chromosomes 3, 17 and 20 in MCF-7).

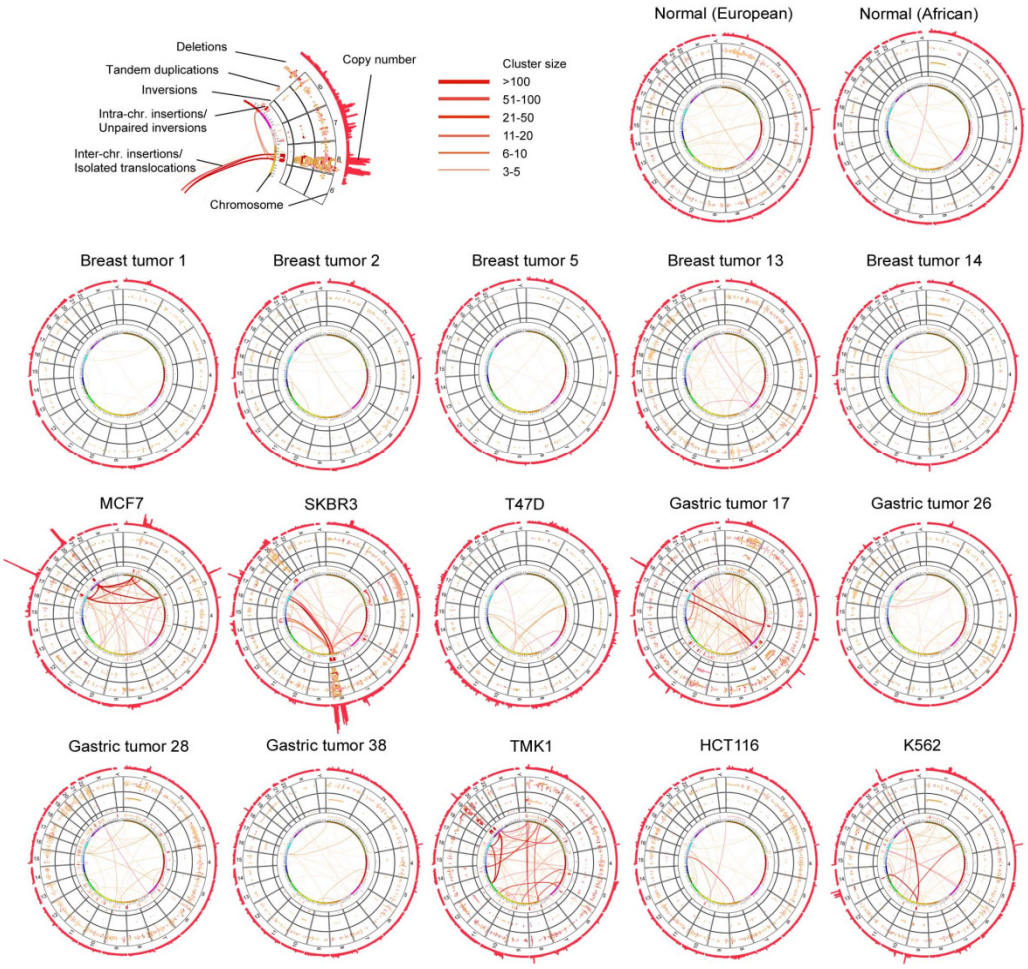

Figure 3.8. Karyo-genomic maps of 15 cancer and 2 normal human genomes.

Genomes are arranged in a circular manner with SV-categories arranged in concentric layers as indicated on the top left. Circular plots have been generated using Circos (Krzywinski et al. 2009).

Copy number analysis by cPET

The whole genome copy number analysis by cPET showed that approximately 0.1% of the normal genomes had an estimated copy number >4, therefore, we defined >4 copies as high copy regions. Regions with an estimated copy number <1 were defined as low copy regions which

included homozygous- and high confidence heterozygous deletions (ca. 0.3% of the normal genomes had <1 copy). We subtracted the regions with high and low copy number, respectively, in the normal genomes (which were likely to be copy number variations [CNVs]) from the cancer genomes (Figure 3.9-3.10). As expected, there were less high copy regions in the five breast and four gastric primary tumors (a total of 1.6 and 12.8 Mb, respectively) compared to breast cancer and gastric cancer cell lines (a total of 86.6 and 86.4 Mb, respectively). SKBR3 displayed five regions on chromosome 8 totaling 11 Mb, and a 12.8 Mb region on chromosome 19 with a copy number just below 1, suggesting a heterozygous deleted state.

We compared high and low copy regions across our eight breast cancer and five gastric cancer genomes (Figure 3.11 and Appendix Tables 5-8). There were 599 loci with an estimated copy number >4 in the breast cancer samples. Four hundred and thirty-five (72.6%) of these regions spanned genes and 101 (24.4%) of the gene-spanning amplified loci were identified in more than one breast cancer sample. We also identified 69 loci with copy number <1 in at least one out of eight breast cancer genomes. Forty-eight (69.6%) of these loci spanned genes and five (10.4%) of the gene spanning low copy loci were identified in more than one genome (Figure 3.11). Similarly, we obtained 24 recurrent high copy and 10 recurrent low copy regions containing at least one gene among the five gastric cancer genomes. It is most likely that cancer-driving genes are among these recurrent high or low copy regions in breast and gastric cancer, respectively.

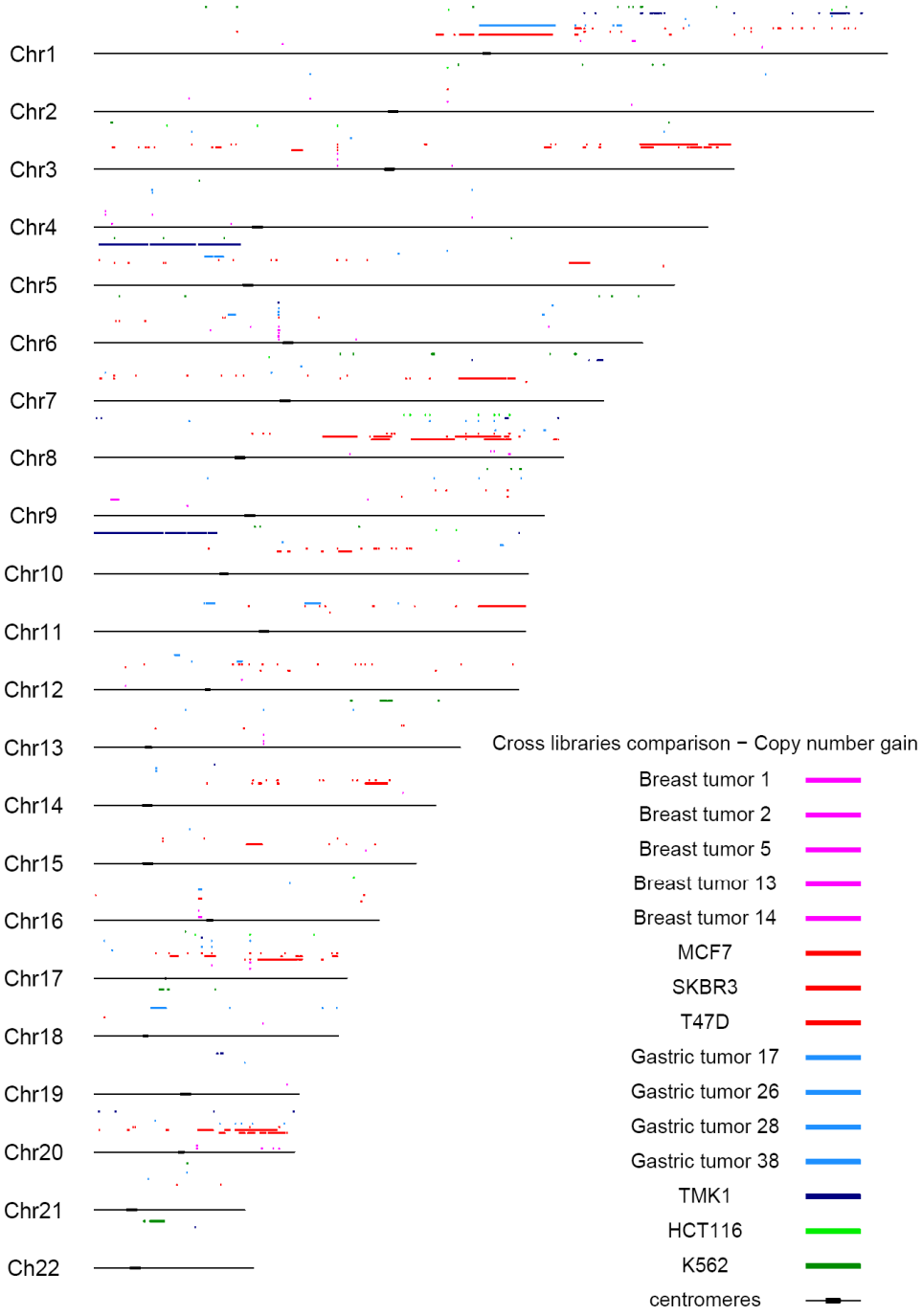

Figure 3.9. Genomic regions of high copy number in 15 cancer genomes.

Chromosomes are represented by horizontal lines and regions with more than two consecutive windows with a copy number >4 are indicated by colored bars. Regions showing amplification in at least one of the two normal genomes which match amplified regions in the cancer genomes by >50% have been excluded.

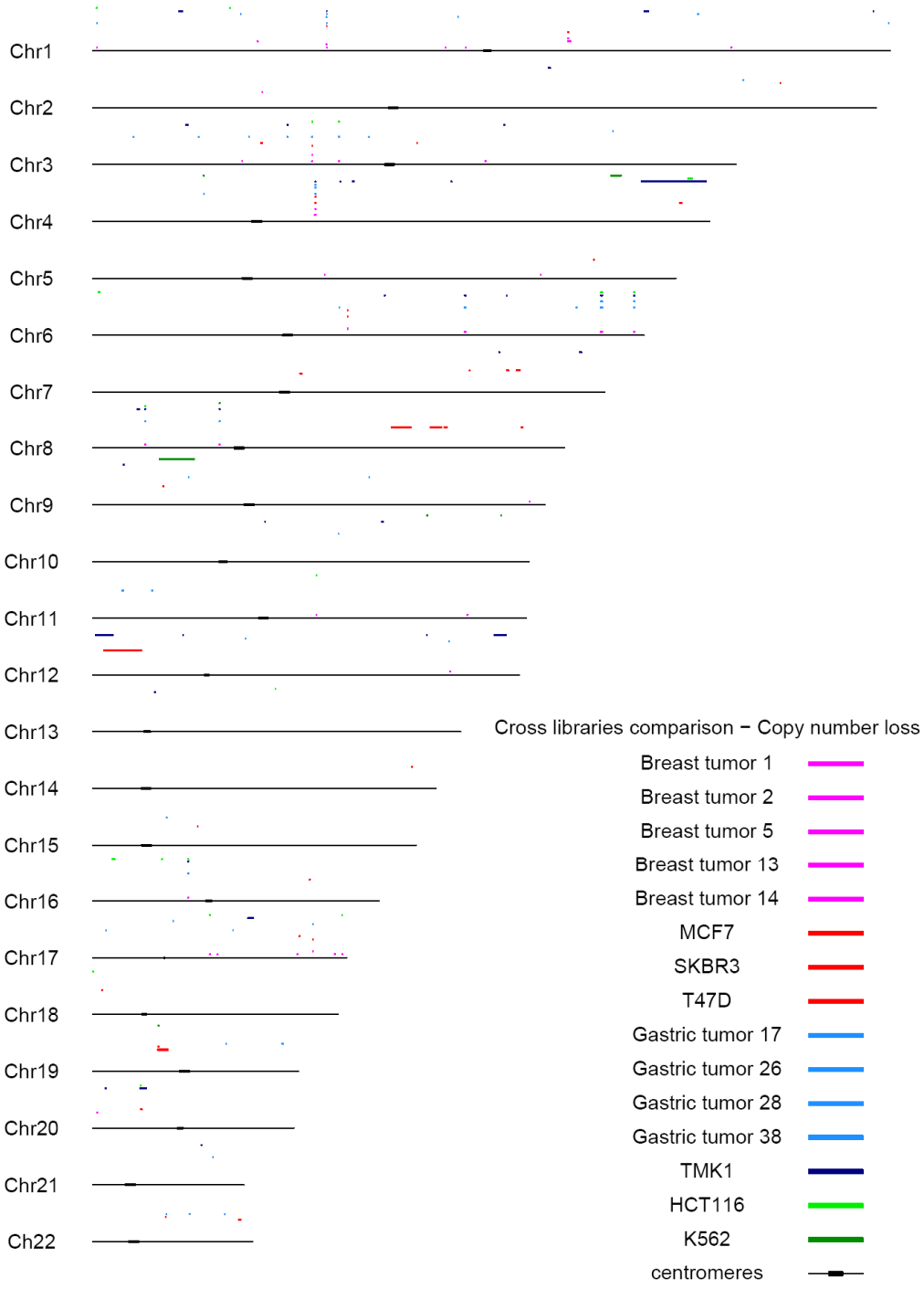

Figure 3.10. Genomic regions of low copy number in 15 cancer genomes.

Chromosomes are represented by horizontal lines and regions with more than two consecutive windows with a copy number <1 are indicated by colored bars. Regions showing low copy number in at least one of the two normal genomes which match the deleted regions in the cancer genomes by >50% have been excluded.

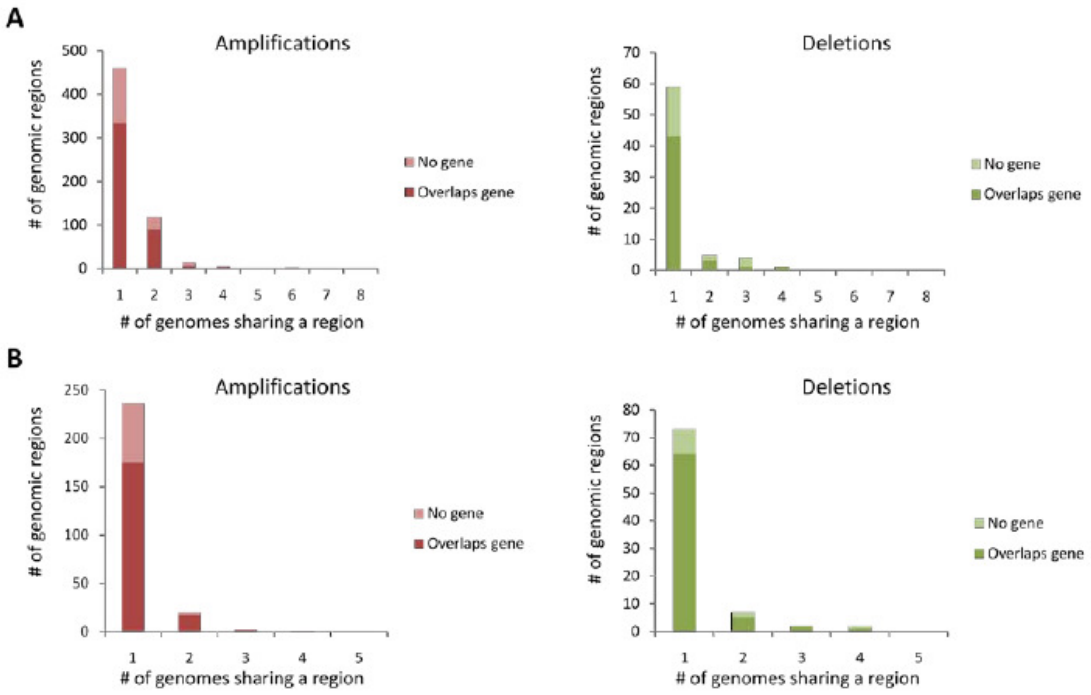

Figure 3.11. Overlap of amplified and deleted regions breast and gastric cancer genomes.

Overlap of amplified and deleted regions in eight breast cancer (A) and five gastric cancer genomes (B). Genomic regions with a predicted copy number >4 are represented in red, regions with a predicted copy number <1 are represented in green. The number of genomes in which the same region was observed is represented on the x-axis. High/low copy regions which matched high/low copy regions observed in at least one of the two normal genomes were excluded.

Copy number support for deletions and tandem duplications

As tandem duplications and deletions (if not reciprocal) are expected to result in copy number changes, we tested this by using the sequence tag based copy number approach with a copy number change of ≥ 0.5 in the expected directions at the predicted breakpoints. Using these criteria, average 52% of the deletions and 14% of the tandem duplications matched with copy number decrease and increase, respectively (Table 3.4). The relatively low fraction of deletions and tandem duplications with copy number support is partly attributable to the fact that the copy number estimation is based on cPET tags. Windows which overlapped with dPETs were omitted from the copy number analysis, and this affected in particular small deletions and tandem duplications. Small tandem duplications are more sensitive to this phenomenon than deletions, since their dPET anchor regions are inside of the tandem duplication when mapped to the reference (Figure 2.3).

Table 3.4 Deletions and tandem duplications with copy number support in each genome

	Deletion	Tandem duplication
European	265	4
African	107	3
Breast tumor 1	121	0
Breast tumor 2	176	7
Breast tumor 5	58	3
Breast tumor 13	294	4
Breast tumor 14	78	2
MCF-7	201	49
SKBR3	336	28
T47D	165	5
Gastric tumor 17	197	1
Gastric tumor 26	173	9
Gastric tumor 28	257	2
Gastric tumor 38	197	3
TMK1	357	31
HCT116	306	19
K562	305	13

Characteristics of SVs in cancer genomes

The structural maps of the 17 genomes should include germ line and somatic SVs, as well as mapping artifacts and assembly errors in the reference genome sequence. We reasoned that if a particular SV was shared among all 17 genomes, this would most likely represent a rare allele or an assembly error in the reference genome. If an SV was observed in multiple, but not all,

genomes, it would most possibly represent a germ line SV that was accumulated in human populations through evolution history. In contrast, if a SV was unique to one particular cancer genome, then it might be considered as derived somatically. Thus, we conducted comparative analysis of all SVs identified by dPET mapping in the 17 genomes, and found 57 different SVs that were common in at least 16 of the 17 genomes (Figure 3.12 and Appendix Table 9); 1,290 SVs that were shared by multiple genomes (2-15 genomes; Appendix Table 10), indicating potential "germ line" origin; 4,527 SVs which were unique to single genomes (Appendix Table 11), of which 4,489 SVs were found only in cancer genomes and which we considered as, most likely, "somatic" events. The median fraction of uniquely observed SVs in the 15 cancer genomes was 26.3% compared to 7.2% in 16 normal genomes (including DNA-PET data of 14 additional normal genomes; p=4.46 x 10^{-7}; Figure 3.13) suggesting that a large fraction (1-7.2/26.3=72.6%) of unique SVs in cancer genomes was of somatic origin.

A

	Rearrangement points (%)	Events (%)
Deletion	18 (32)	18 (42)
Tandem duplication	1 (2)	1 (2)
Unpaired inversion	3 (5)	3 (7)
Inversion	26 (46)	14 (33)
Intra-chr. insertion	4 (7)	2 (5)
Isolated translocation	1 (2)	1 (2)
Complex inter-chr.	4 (7)	4 (9)
Total	57 (100)	43 (100)

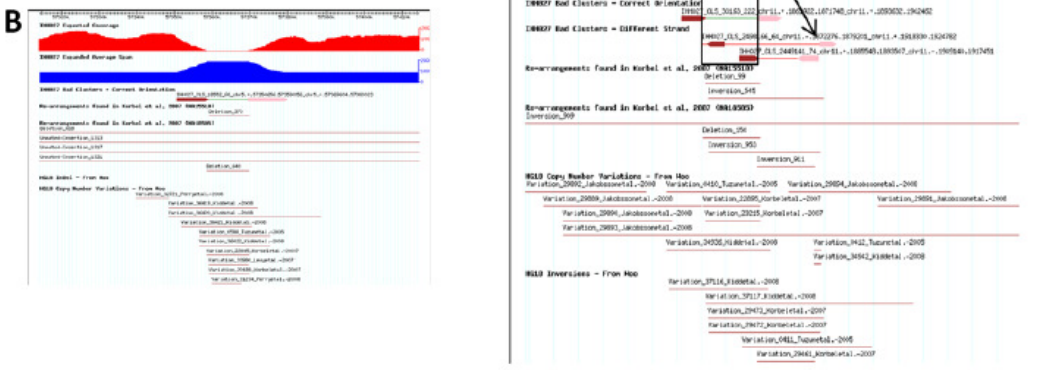

Figure 3.12. Predicted SVs that were observed in 16 or 17 out of 17 genomes by dPET clusters.

(A) SV categories of the 57 dPET clusters observed in 16 or 17 genomes. Deletions and inversions are the predominant SV categories. (B) Example of a deletion. The predicted deletion matches entries reported by Korbel et al. (Korbel et al. 2007) downloaded from the Database of Genomic Variants (DGV, Iafrate et al. 2004) by 94%. Red track represents the coverage of the cPETs, blue track represents the average span between the two tags of a PET. Dark red and pink arrows connected by a green line represent the left and right anchor region, respectively, flanking the deletion. Horizontal lines (bottom) represent SV entries in DGV. (C) Example of an inversion. Black box indicates inverted segment. The cPET coverage drops down to zero at the predicted breakpoints, suggesting a homozygous genotype. (D) Intra-chromosomal inverted insertion of 21 Kb on chromosome 11. The combination of a deletion with an inverted insertion demonstrates the cut (black box) and paste (black arrow) nature of the deletion. Previously reported deletions and inversions are likely to describe the same structure.

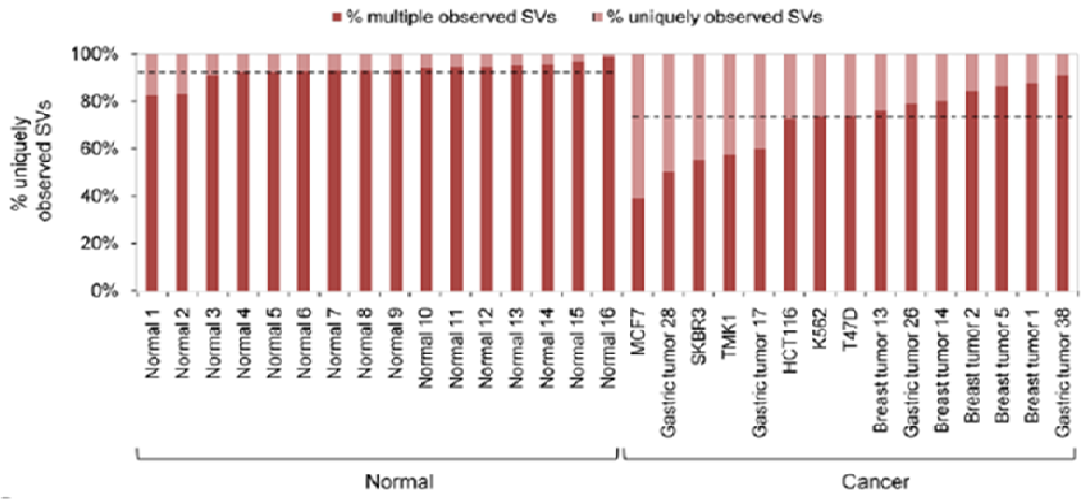

Figure 3.13. Unique vs. multiple observations of SVs identified by DNA-PET sequencing.

SVs of 2 normals of this study (Normal 15 and 16), 14 additional normals (unpublished data) and 15 cancer genomes (x-axis) were compared with each other and with the published SVs of 10 normal genomes (Kidd et al. 2008; Korbel et al. 2007). SVs that were observed only in one genome (unique) are shown in light red and SVs observed in two or more genomes are in dark red. Proportions of uniquely and multiple observed SVs are given on the y-axis. Dashed lines indicate median of uniquely observed SVs in normal and cancer genomes, respectively.

The PCR validation rate for multiply observed SVs was higher than the rate for uniquely observed SVs (78.4% vs. 69.2%, respectively), indicating a higher false discovery rate for the unique category. Among the 17 genomes, we found 62 and 96 unique SVs, respectively, in the normal genomes which could represent novel private germ line SVs. We also found an average unique SVs of 115 in breast tumors (range: 43-306), 344 in gastric tumors (range: 73-669), 428 in breast cancer cell lines (range: 104-651), and 584 in the single gastric cancer cell line, TMK1. Although the comparison of the European and African normal samples with European (breast tumors) and East-Asian (gastric tumor) cancer samples is not straightforward, the increase of unique SVs in primary tumors and cell lines can be explained by somatic rearrangements. Some

SV classes appear to be more likely germ-line variants than others. Most inversions and intra-chromosomal insertions found in the 17 genomes were highly shared among multiple genomes and were significantly represented in the two normal genomes, suggesting that the majority of SVs in these two categories are most likely of “germ line” origin (Figure. 3.14). Most of the isolated deletions and inter-chromosomal insertions could also be considered as “germ line” SVs. In contrast, tandem duplications, isolated translocations, unpaired inversions, and complex rearrangements were over-represented in genome-unique “somatic” SVs (Figure. 3.14).

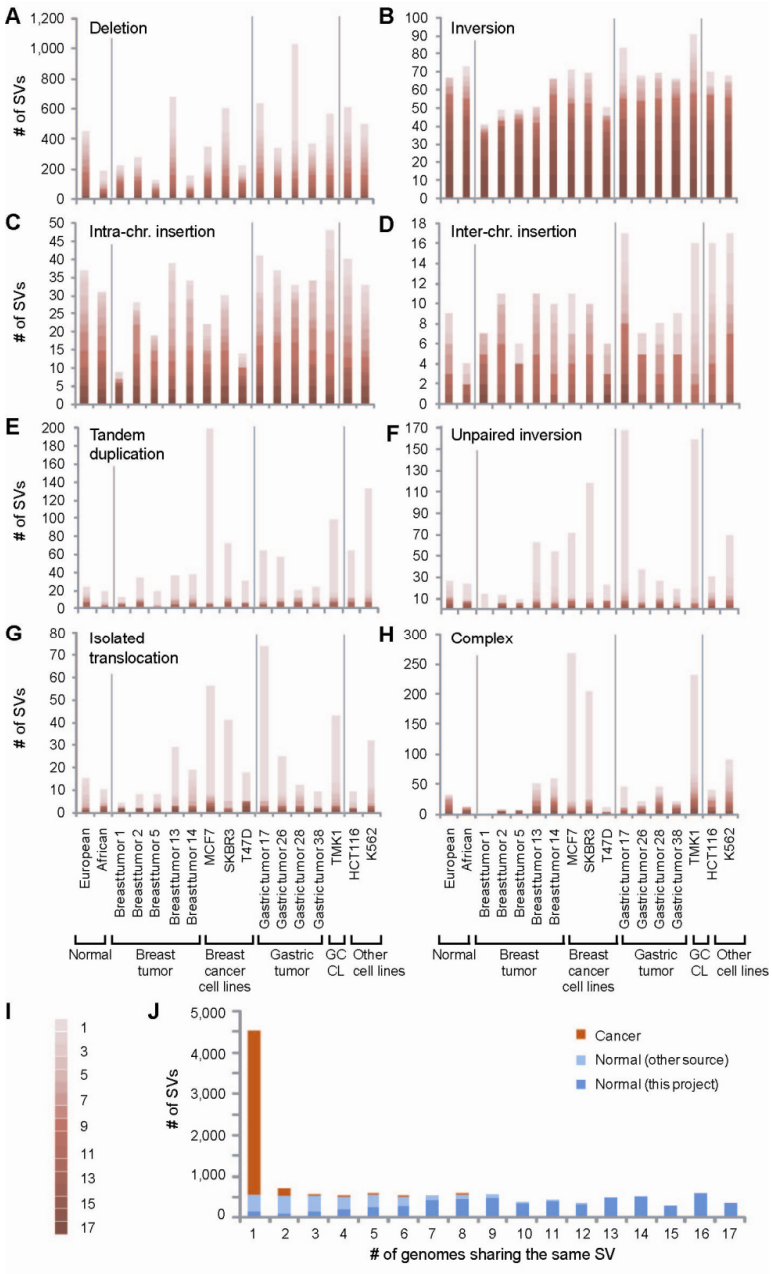

Figure 3.14. Comparison of SVs across 15 cancer and 2 normal genomes.

(A–H) Frequencies (y-axis) of the indicated SV categories are shown for the individual genomes (*x-axis*). Cancer groups are separated by vertical gray lines. Degree of recurrent observation of the same SV is indicated in (*I*) where 1 represents the observation in one genome and 17 represents the observation in all 17 genomes. (*J*) SVs which were observed in the normal individual(s) or which were observed in the cancer genomes but match those observed in the normal individuals or match by >80% earlier described events (Kidd et al. 2008; Korbel et al. 2007) are indicated in dark blue. SVs which were also observed in other 14 normal individuals are indicated in light blue. SVs observed only in cancer genomes are indicated in orange. The *x-axis* represents the number of genomes which share a particular SV and *y-axis* represents the frequency.

As no DNA samples of paired non-cancer (normal) tissues were available for the 15 cancer genomes, we used the SVs identified in 12 unrelated normal individuals, two of this study and 10 published previously (Kidd et al. 2008; Korbel et al. 2007), to filter and thereby strongly enrich the set of cancer SVs for somatic events. Since the aim of this project was to analyze the general characteristics of genome structural changes in breast and gastric cancer, the strong enrichment for somatic events was considered sufficient. Using the common SV filtering approach, we classified the SVs of the 15 cancer genomes that were sheared with normal genomes in this study or in previous reports (Kidd et al. 2008; Korbel et al. 2007) as "normal genome SVs" (n = 5,105, Figure 3.15), and the SVs found only in at least one of the 15 cancer genomes but not represented in normal genomes as "cancer genome SVs" (n = 6,410, Figure 3.15). Most cancer SVs (87.3%) were identified only in one genome, whereas normal SVs were mostly shared by multiple genomes (Figure 3.16). This suggests that most cancer specific SVs are likely to be private mutations.

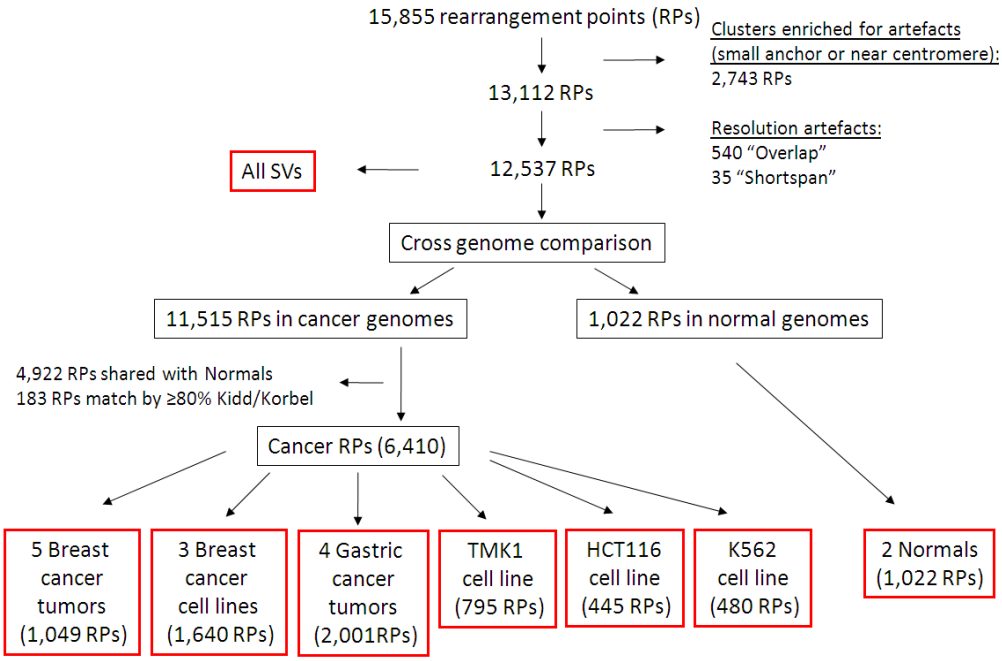

Figure 3.15. Flow chart of dPET cluster data.

The rearrangement points include redundant observations in different genomes.

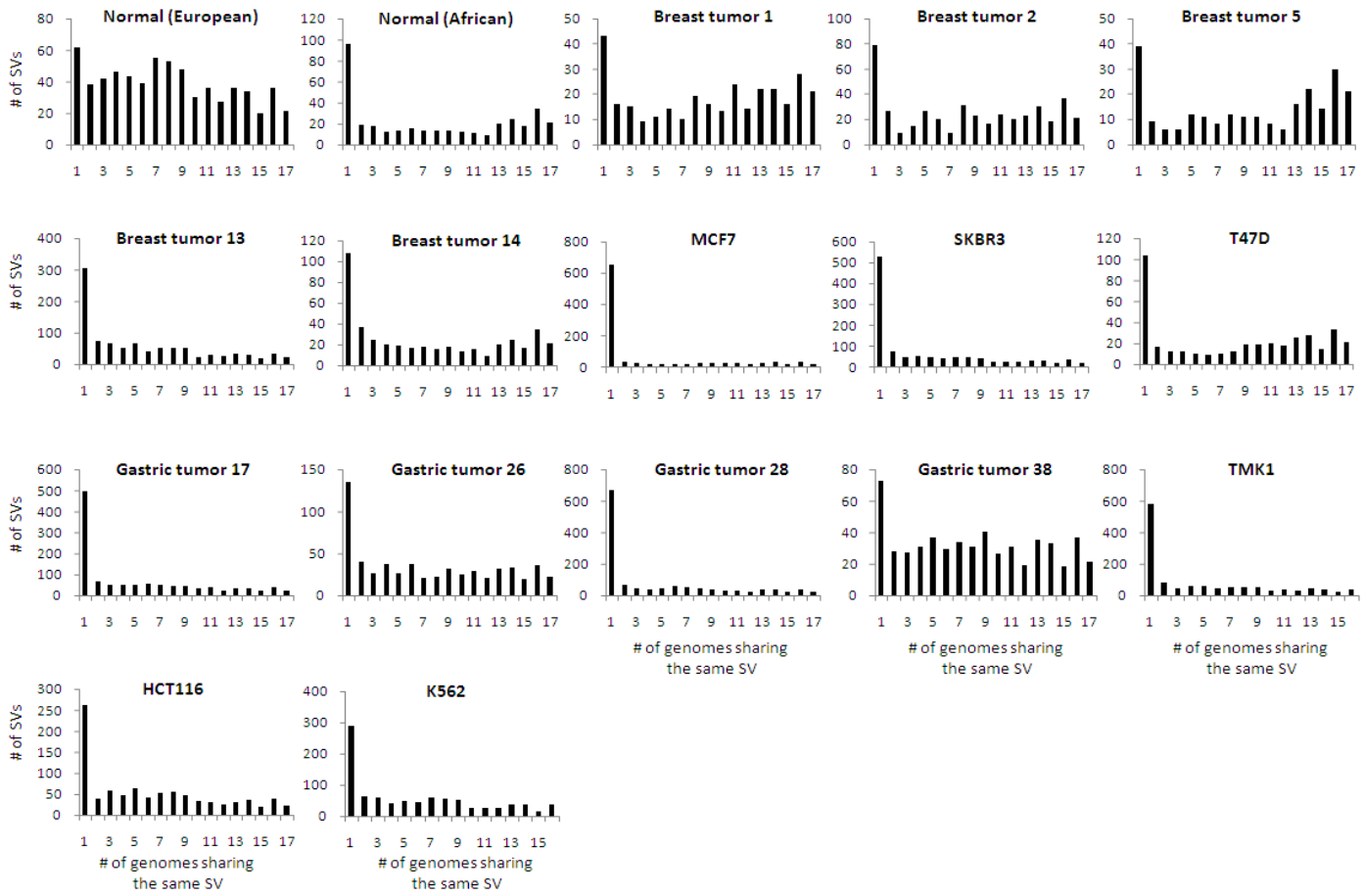

Figure 3.16. Numbers of observed SVs across 17 human genomes analyzed by PET sequencing.

Note the high proportion of uniquely observed SVs in the cancer genomes compared to the normal European. The normal African genome also shows a high proportion of uniquely observed SVs. This may be due to differences in ethnicity since the breast cancer tumors and cell lines are derived from European and gastric cancer tumors from Asian individuals.

To evaluate the efficiency of our common SV filtering approach, we analyzed the DNA-PET data of a tumor/blood pair and compared our filtering approach with the matched pair filtering. Of 1,144 SVs identified in the tumor, 716 were classified as tumor specific SVs using the common SV filtering approach whereof 236 were detected in the paired blood sample of the tumor indicating their germ line origin (Figure. 3.17). This resulted in 67% (480/716) of the SVs which were assigned as tumor specific by the common filtering that were also assigned as tumor specific using the DNA-PET information of the paired blood sample. The cancer SVs were therefore strongly enriched for somatic events.

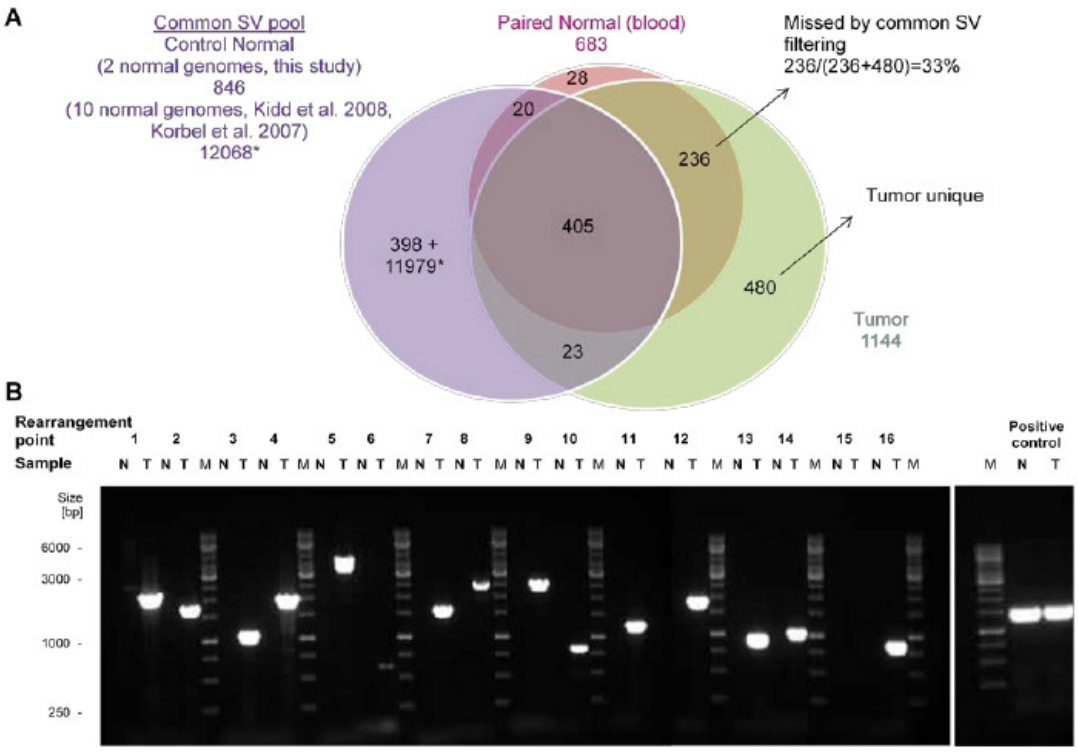

Figure 3.17. Comparison of germ line variation filtering by a paired sample approach vs. the common SVs approach and validation of somatic SVs.

(A) SVs of a tumor (green) and the matched (normal) blood (pink) were intersected with each other and with SVs identified in the two normal individuals of this study and published SVs of ten normal individuals (Kidd et al. 2008; Korbel et al. 2007), respectively (purple). The numbers of SVs are indicated for each category and section of the diagram. Seven hundred and sixteen SVs (236+480) were classified as cancer SVs by the common SV filtering of which 236 SVs (33%) were of germ line origin based on the paired sample. This fraction further drops by increasing the pool of normal samples. Four hundred eighty SVs were identified as somatic mutations by the combined filtering approach, includes redundant SVs in different individuals. (B) PCR validation of predicted somatic SVs in (A). Primer pairs flanking 16 somatic rearrangement points predicted by DNA-PET were used to PCR amplify potential rearrangement points with DNA templates from the normal blood (N) and tumor (T), respectively. PCR products were separated together with a molecular size marker (M) by electrophoresis in 1% agarose gels and stained by ethidium bromide. Reactions of 15 out of 16 predicted somatic rearrangement points resulted in specific PCR amplicons in the tumor but not in normal samples. The positive control PCR reactions from a genomic location of a germ line SV show comparable input DNA templates and PCR efficiency for the tumor and normal samples.

Although the number of breast and gastric cancer samples in our study was limited, we attempted to identify potential recurrent rearrangements. In the comparative analysis, we observed 244 cancer SVs in more than one breast cancer and 256 cancer SVs in more than one gastric cancer genome. After further filtering by additional SVs of 14 normal individuals mapped by DNA-PET (unpublished data) and overlap (>50%) with entries in the Database of Genomic Variants (DGV) (Iafrate et al. 2004), we identified 15 potentially recurrent SVs in breast cancers, and 21 in gastric cancers (Table 3.5). Since we expect different breakpoints as a genuine criterion for a recurrent event in two different genomes, we analyzed the breakpoints by PCR and Sanger sequencing. The majority of the potentially recurrent events showed exactly the same breakpoints or could not be validated (27/36). The remaining events, four in breast cancer and five in gastric cancer (Appendix Table 12), had highly homologous sequence features in the breakpoint junction regions, so that the exact breakpoint positions could not be precisely determined by this method. Among them might be some real recurrent events. All except one are intra-chromosomal and six out of nine are alterations shorter than 25 kb. Many of them have the potential to alter gene functions, either in coding sequences (deletion of exons or altering introns in 3/9) or in regulatory regions (3/9). However, none of the putative recurrent SVs represent putative fusion transcripts. Further investigation of the breakpoint characteristics and analysis in larger cohorts of tumor specimens and in normal DNA will be necessary to distinguish recurrent cancer SVs from low frequency normal SVs.

Table 3.5 Effect of *in silico* filtering of potentially recurrent breast and gastric cancer breakpoints: most cancer breakpoints are uniquely observed

Cancer	Type	Total percent	% of remaining breakpoints after filtering by	
			Additional normal genomes[1)]	DGV[2)]
Breast	Uniquely observed	100 (n=2,078)	83.7 (n=1,739)	80.8 (n=1,680)
	Potentially recurrent	100 (n=244)	15.2 (n=37)	6.1 (n=15)
Gastric	Uniquely observed	100 (n=2,135)	84.1 (n=1,795)	80.7 (n=1,723)
	Potentially recurrent	100 (n=256)	8.6 (n=22)	8.2 (n=21)

[1)] 14 additional normal genomes analyzed by paired-end sequencing (unpublished data)
[2)] Entries in the Database of Genomic Variants which match by ≥50%

We further compared the breast cancer SVs identified in this study with SVs identified in the 24 breast cancer genomes sequencing paper (Stephens et al. 2009).Twenty different SVs overlapped by >50% with SVs identified in our study (Table 3.6). Most of them were large rearrangements on chromosome 8 in MCF-7 and SKBR3, respectively. A homozygous deletion of 170 kb on chromosome 9 (21,809,532-21,979,622) in MCF-7, which deletes the tumor suppressor gene *CDKN2A*, overlaps by 70% with a deletion reported by Stephens and colleagues (Stephens et al. 2009). *CDKN2A* is a tumor suppressor gene and homozygous or heterozygous deletions of *CDKN2A* in breast cancer and other cancer had been reported earlier (Jonsson et al. 2007) which confirm the relevance of this locus for breast cancer.

Table 3.6 Breast cancer specific rearrangements which overlap by ≥50% with events reported by Stephens et al. (2009)

Genome	SV type	Coordinates	Cluster size	Span	Truncated genes	Genes spanning SV	# genes inside SV	SV type Stephens et al.	Overlap [%][1)]
MCF-7	Deletion	chr7:69635775..69861732	18	255,957		*AUTS2*	0	Deletion	68
SKBR3	Deletion	chr8:89529429..425352316	4	35,822,887			133	Amplified	57
SKBR3	Deletion	chr8:92619012..109850006	132	17,230,994	*TMEM74*		76	Deletion	68
MCF-7	Deletion	chr8:106854866..124772942	18	17,918,076	*ZFPM2,ANXA13*		49	Amplified	85
SKBR3	Complex intra-chr.	chr8:108941884..133280776	15	24,338,892	*KCNQ3*		n.a.	Amplified	64
SKBR3	Deletion	chr8:110109746..112776734	66	2,666,988			14	Amplified	57
MCF-7	Deletion	chr8:112436691..121631999	46	9,195,308	*SNTB1*		21	Amplified	63
SKBR3	Tandem duplication	chr8:89760331..112880172	7	23,119,841			95	Amplified	53
SKBR3	Deletion	chr8:114510462..121971682	70	7,461,220	*CSMD3*		21	Amplified	51
SKBR3	Tandem duplication	chr8:91601813..114532206	25	22,930,393			91	Amplified	53
SKBR3	Deletion	chr8:115867498..127928991	3	12,061,493			50	Amplified	77
SKBR3	Deletion	chr8:115872706..123915028	5	8,042,322	*ZHX2*		23	Amplified	84
SKBR3	Complex intra-chr.	chr8:120588540..142127407	13	21,538,867			n.a.	Amplified	60
SKBR3	Deletion	chr8:120938433..141865475	21	20,927,042	*PTK2*		67	Amplified	62
SKBR3	Tandem duplication	chr8:88485220..123034720	8	34,549,500			121	Amplified	59
SKBR3	Complex intra-chr.	chr8:118044568..124227138	9	6,182,570	*WDR67,SLC30A8*		n.a.	Amplified	64
SKBR3	Complex intra-chr.	chr8:128999924..132263367	16	3,263,443	*PVT1*		n.a.	Amplified	50
MCF-7	Deletion	chr8:129972165..132903766	48	2,931,601			5	Amplified	63
MCF-7	Deletion	chr9:21809087..21979717	14	170,630	*MTAP,CDKN2A*		2	Deletion	70
MCF-7	Tandem duplication	chr18:41563696..41659356	31	95,660	*SLC14A1*		0	Tandem duplication	73

[1)] % of overlap between events identified in this study and in the study by Stephens et al. (2009)

We also investigated whether the segmental size of SVs involved in cancer genomes has some specific characteristics. Interestingly, the sizes of tandem duplications of breast and gastric cancers were clearly larger than those found in normal genomes (median tandem duplication size in normals =22 kb vs. 100 kb in breast cancer cell lines [p=8x10^{-6}], 143 kb in gastric tumors [p=4.5x10^{-4}], and 91 kb in TMK1 [p=5x10^{-4}]; Figure. 3.18). The size of unpaired inversions was also larger than that of other SV categories. Compared to normal genomes, the size of unpaired inversions was significantly larger in breast cancer cell lines (p=0.007). Inversions and insertions did not show significant size differences between cancer and normal.

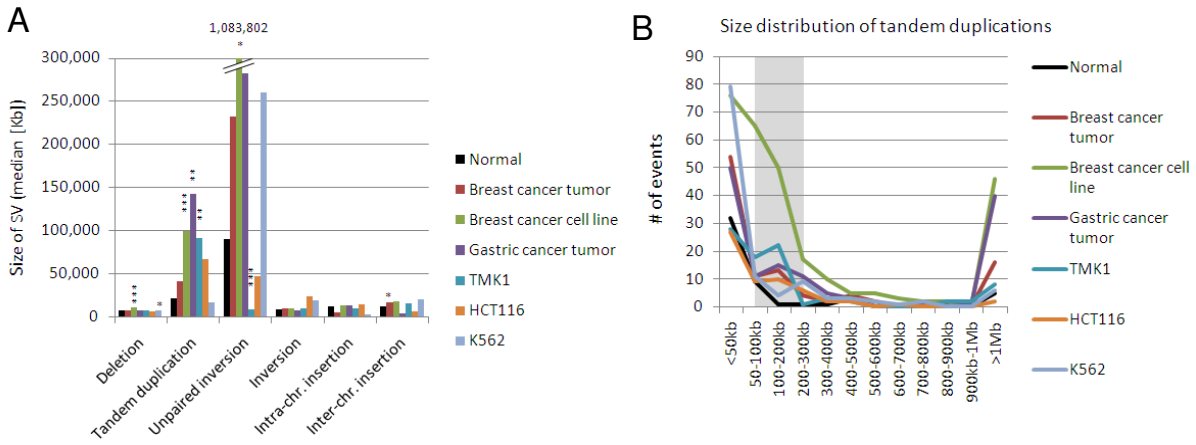

Figure 3.18. Distance of breakpoints.

(A) Median sizes of different SV categories. SVs were grouped into seven categories: two normal individuals, five breast cancer tumors, three breast cancer cell lines, four gastric cancer tumors, TMK1, HCT116 and K562. Of all cancer data sets, SVs which matched those observed in the normal individuals were excluded. Mann-Whitney *U* test P values <0.01 compared to normal SV sizes are displayed: *, P<0.01; **, P<0.001; ***, P<0.0001. (B) Size distribution of tandem duplications indicates an overrepresentation of size 100-300 Kb (shaded in gray) in the cancer categories relative to normal. SVs were grouped as described in (A).

Characteristics of breakpoints

We were interested in understanding whether SVs of potential somatic origin in cancer genomes (identified only in cancer after subtracting normal SVs) have distinctive features around the break points that would provide insights for the underlying mechanisms of cancer genome rearrangements. The set of 12,537 SVs identified in this study has a median resolution of <400 bp, based on the comparison of the DNA-PET predicted breakpoint coordinates with PCR and Sanger sequencing determined breakpoint coordinates. One thousand and twenty-two of the SVs were likely to be normal variants, and 6,410 were likely to be cancer associated (Figure. 3.15). This pool of precise SVs represents an unprecedented resource for the determination of sequence features associated with rearrangement events.

Overall, we have observed that significantly more normal SVs (from germ line DNA) had breakpoint sequence homology than cancer SVs, with striking differences for tandem duplications, unpaired inversions, and complex rearrangements: ~60% of the normal category had breakpoint homology as compared to ~20% of the cancer category (Figure 3.19.A, $p<10^{-15}$). This result suggests that non-homology-based rearrangements are characteristic for cancer genomes, which is in accordance with the understanding that a significant proportion of normal SVs is mediated by nonallelic homologous recombination (NAHR), whereas the majority of somatic events in rearranged cancer genomes is based on non-homologous end-joining (NHEJ) (Hampton et al. 2009; Raphael et al. 2008). At the level of specific SV types, among the normal SVs, deletions showed the lowest fraction of breakpoint sequence homology (24.9%) and inter-

chromosomal insertions showed the highest fraction (90.9%).

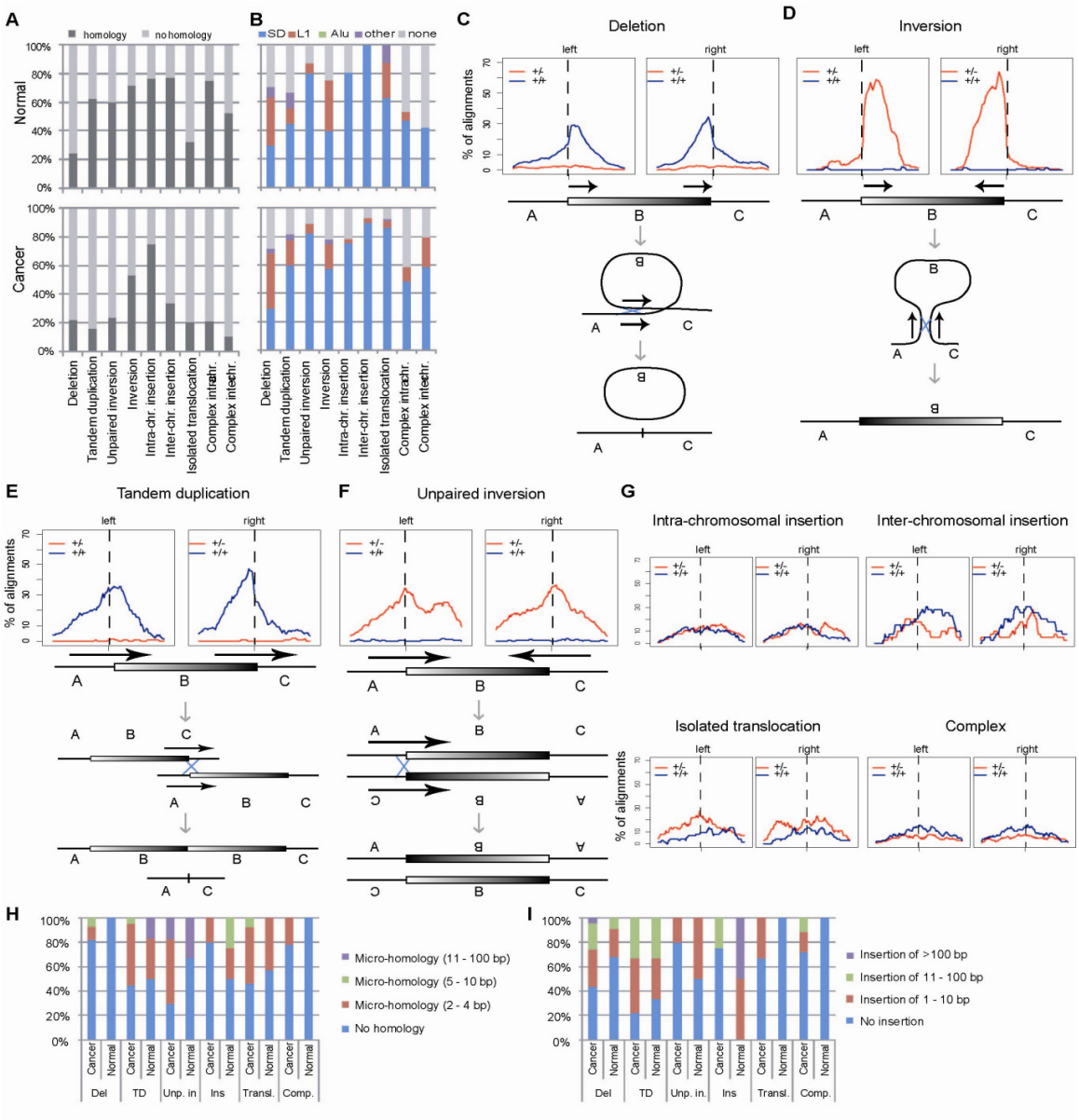

Figure 3.19. Sequence features of rearrangement points.

(A) Normal and cancer SVs were stratified for the presence or absence of sequence homology between the two breakpoints of a pair. (B) Sequence features for SVs with sequence homology. SD, segmental duplication; L1, long interspersed nuclear element 1; Alu, Alu element; other, other UCSC annotated multi copy sequences. (C-G) Cumulative alignment of sequence homologies at predicted breakpoints (vertical dashed lines). Blue lines indicate sequence homologies between the same strand of genomic DNA, red lines indicate sequence homologies between different strands. (C-F) Middle, schematic representations of genomic regions. Boxes indicate regions between breakpoints; arrows indicate orientation of sequence homology. Bottom, schematic representation of recombination mechanisms. Blue "X" indicates locus of recombination. (H) Micro-homology of sequenced breakpoint pairs. (I) Insertions of DNA sequences of unknown origin for breakpoint pairs without micro-homology.

Segmental duplications were the major source for sequence homology of all SV types and account for all homologies of normal inter-chromosomal insertions (Figure 3.19.B). In cancer SVs, tandem duplications and complex rearrangements had the least homology, followed by deletions and isolated translocations.

We further investigated the location and orientation of the homologous breakpoint sequences, and found that the homologous sequences of deletions and tandem duplications were positioned in tandem orientation (head to tail or on the same strand), whereas for inversions and unpaired inversions, the two homologous sequences were positioned in reverse orientation (head-to-head or on different strands, Figure 3.19. C-F). The specific orientation of homology, which is in accordance with the assumed DNA looping structures, demonstrates that the underlying mechanism is NAHR (reviewed in (Gu et al. 2008). Interestingly, we observed increased frequency of sequence homology at the predicted breakpoints for deletions (n=910) and in particular for inversions (n=160) with a higher frequency of sequence similarity at the predicted breakpoint inwards of the anchor regions (Figure 3.19. C-D). In contrast, the homologous sequences at the breakpoints of unpaired inversions (n=203) showed relatively even distributions on both sides of the breakpoints, indicating that the breakpoint/homology correlation is less specific (Figure 3.19. F). These deposition patterns of homologous sequences suggest a distinctive DNA alignment mechanism for inversions as compared to unpaired inversions. We also observed a moderate difference in alignment profiles between deletions and tandem duplications (n=149). Other SVs such as insertions (n=194), isolated translocations (n=73), and complex events (n=172) showed less characteristic patterns, suggesting other underlying mechanisms (Figure 3.19. G).

For SVs that had no significant homology (BLAST score <300), we conducted detailed analysis of the junctions by PCR and sequencing to investigate the frequency of possible micro-homology, a reported signature of NHEJ (Cahill et al. 2006). We did not observe general differences in the micro-homology rates between cancer and normal SVs (Figure 3.19. H-I and Appendix Table 13). Deletions (cancer and normal) showed the lowest degree of micro-homology which might suggest a mechanism without open DNA-end intermediates.

The impact of SVs on genes

We investigated whether more gene structures were altered in the cancer genomes compared to the two normal genomes. Since the majority of the SVs which were enriched in the cancer genomes were uniquely observed (Figure 3.14. J), we compared the numbers of genes which were affected by genome-unique breakpoints. Two breast tumors, two gastric tumors and all cell lines showed a larger number of genes which were predicted to be fused compared to the two normal genomes (Figure 3.20. A), with MCF-7 showing the largest number of predicted fusion genes (n=52). A similar pattern was observed for truncated genes with an additional gastric tumor showing a high number of affected genes (Figure 3.20. B). We observed the largest number of genes with deletions inside their gene bodies for gastric tumor 28 followed by breast tumor 13, which correlates with the largest numbers of small deletions (<4 Kb) in these genomes (324 and 244, respectively; Figure 3.20. C). The number of genes in amplified regions (copy number >4) was higher for all cell lines except HCT116 compared to the two normal genomes, with SKBR3 having the largest number of 845 genes (Figure 3.20. D). The data reflect a quantitative difference of genes affected by structural changes for most of the sufficiently sequenced cancer genomes with the same trend in breast and gastric cancer. Cell lines show more genes which are fused, truncated or in amplifications compared to primary tumors. A

detailed analysis of the impact of the genomic rearrangements on the transcriptome of the eight breast cancer samples is provided elsewhere (Inaki et al. 2011).

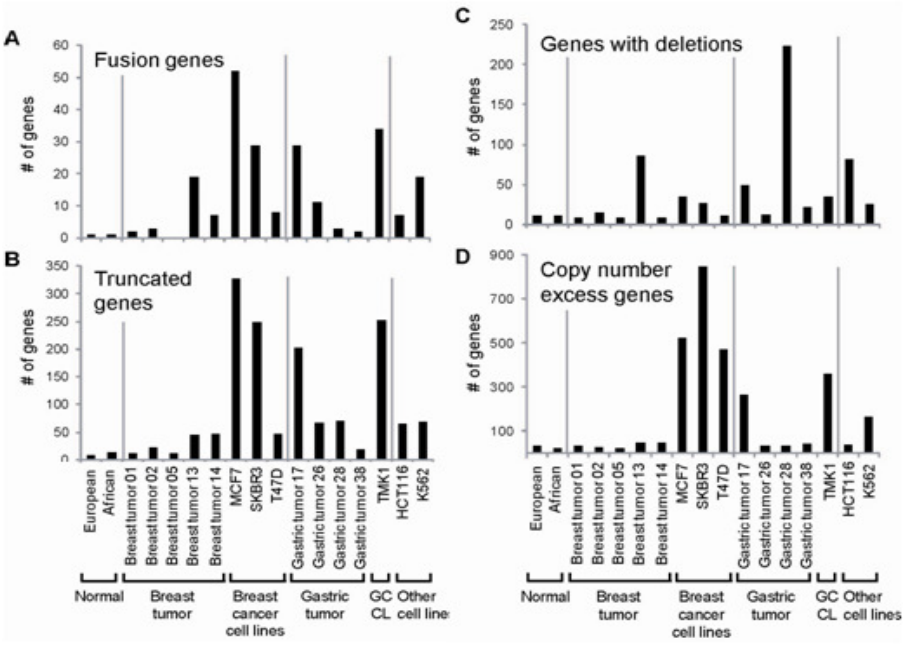

Figure 3.20 Genes affected by SVs.

Number of fusion genes (A), truncated genes (B), and genes with deletions inside their gene bodies (C) which are predicted by uniquely observed breakpoints. (D) Genes in regions with predicted copy numbers >4.

Whereas breakpoints in normal genomes occurred as frequently in gene deserts as in other regions, cancer breakpoints were significantly under-represented in gene deserts compared to expectation ($p<10^{-15}$; Figure 3.21). Intriguingly, we found that cancer breakpoints were not enriched within gene bodies but within 10 kb up- and downstream of genes ($p<10^{-15}$). Taken together, these data suggest that perturbation of gene regulation may be under positive selection in the evolution of a cancer cell.

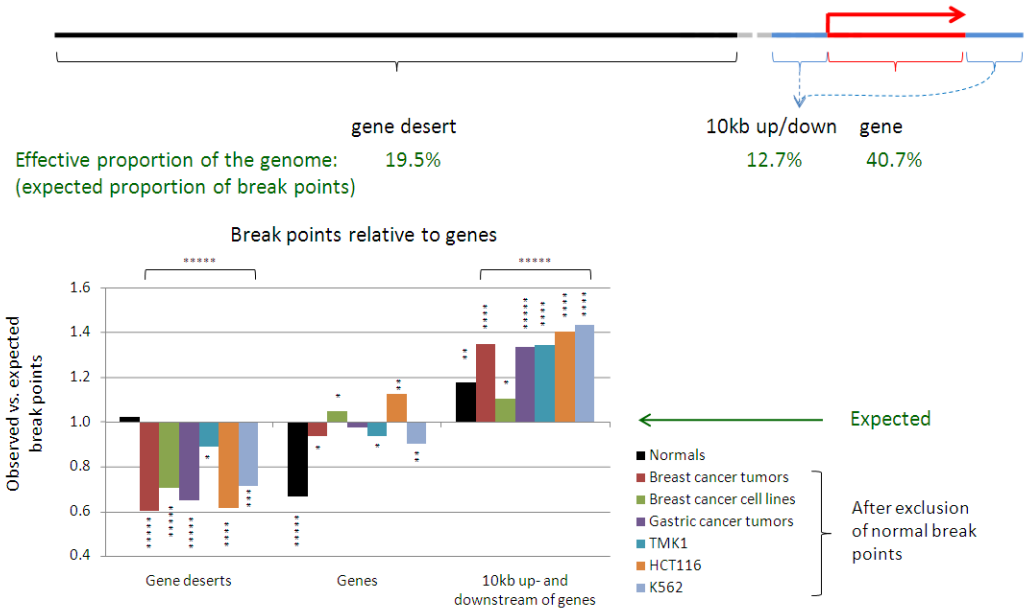

Figure 3.21. Observation of breakpoints relative to genes.

Breakpoints were split in seven categories: normal individuals (n=2), breast cancer tumors (n=5), breast cancer cell lines (n=3), gastric cancer tumors (n=4), and the cell lines TMK1, HCT116, and K562. Of all cancer data sets, breakpoints which matched SVs observed in the normal individuals were excluded. Gene deserts were defined as 1 Mb regions without gene annotation (RefSeq +10 kb). The proportion of gene deserts vs. the rest of the genome was taken as the expected proportion of breakpoints and compared with the observed number of breakpoints. Similarly, 10 kb regions up- and downstream of gene and RefSeq genes, respectively, were analyzed. *, $P<0.05$; **, $P<0.01$; ***, $P<10^{-4}$; ****, $P<10^{-5}$; *****, $P<10^{-10}$ Chi^2 test. Statistics for individual genomes show the same trend.

Small deletions, tandem duplications and insertions within genes can affect transcripts by deleting, duplicating or inserting exons or altering splicing. In particular, we have validated on the genomic and transcription level in MCF-7 a 53.6 kb deletion in *ITPR1* resulting in an in-frame deletion of exons 42-57 (Figure 3.22). Partial deletions and a missense mutation of the inositol triphosphate receptor type I gene (ITPR1) have been reported to be responsible for Spinocerebellar ataxia type15 (SCA15), a pure ataxia characterized by very slow progression.

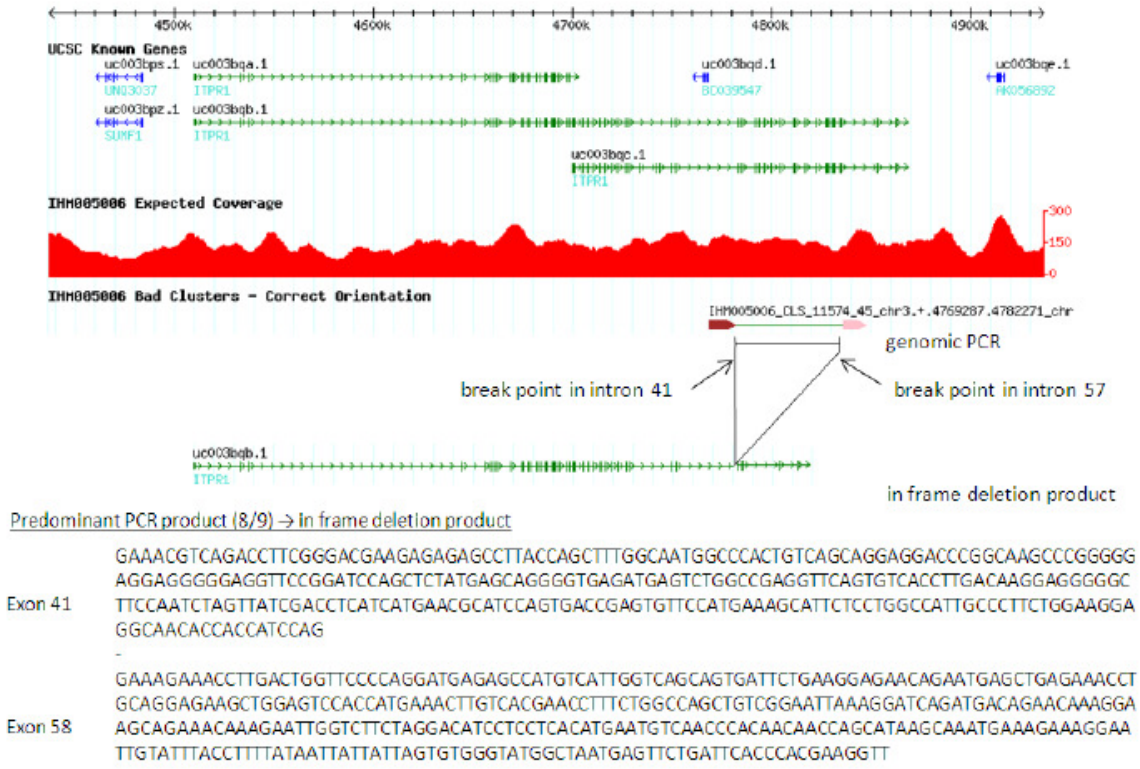

Figure 3.22. Deletion of 15 exons of *ITPR1* in MCF-7.

Genome Browser screen shot of the *ITPR1* locus in MCF-7 (top) shows cPET coverage track (red) and a deletion (dark red and pink arrows). Genomic breakpoints have been validated by PCR and Sanger sequencing. Gene structure derived from breakpoints and RT-PCR is shown below. Bottom: RT-PCR and sequencing result confirms the deletion of exons 42-57.

Gene ontology (GO) analysis of all genes with breakpoints showed cell adhesion-mediated signaling and cell adhesion as the two most significantly overrepresented gene categories in breast and gastric cancer (Table 3.7). This is in accordance with the understanding that epithelial cancers frequently show aberrations in cell-cell and cell-extracellular matrix interactions.

Table 3.7 Gene ontology (GO) analysis of genes[1)] with breakpoints in breast and gastric cancer genomes

Breast cancer						Gastric cancer					
Biological Process[2)]	RefGene	Expected	Observed	+/-	P value		RefGene	Expected	Observed	+/-	P value
Cell adhesion-mediated signaling	386	20.88	51	+	2.22E-06	Cell adhesion	592	31.61	73	+	2.62E-09
Cell adhesion	592	32.02	65	+	3.47E-06	Cell adhesion-mediated signaling	386	20.61	49	+	9.66E-06
Neuronal activities	561	30.34	54	+	1.52E-03	Signal transduction	3256	173.84	222	+	1.53E-03
Biological process unclassified	5972	322.97	269	-	3.04E-03	Synaptic transmission	275	14.68	33	+	3.24E-03
Electron transport	230	12.44	2	-	1.05E-02	Cell communication	1207	64.44	98	+	4.47E-03
Other intracellular signaling cascade	212	11.47	27	+	1.12E-02	Neuronal activities	561	29.95	51	+	6.94E-03
Cell communication	1207	65.28	97	+	1.16E-02	Biological process unclassified	5972	318.84	273	-	2.36E-02
Chemosensory perception	204	11.03	1	-	2.67E-02	Chemosensory perception	204	10.89	1	-	3.04E-02
Developmental processes	2065	111.68	144	+	2.97E-02						

[1)] Genes with breakpoints are all genes which have a breakpoint in at least one out of eight breast cancer and one out of five gastric cancer genomes, respectively. Breakpoints which matched those observed in the normal individuals were excluded.

[2)] GO analysis was performed using Panther (www.pantherdb.org) with RefSeq genes downloaded from UCSC (http://genome.ucsc.edu/) as reference. “Unclassified” biological process (n=1 for each list of biological processes) has been excluded.

Genomic architecture of amplified regions in MCF-7

A hallmark of cancer genomes is the complex amplification of DNA segments (Hicks et al. 2006; Jonsson et al. 2007). This is evident in the 15 structural maps of cancer genomes (Figure. 3.6). Indeed, amplifications and inter-chromosomal translocations were observed in both primary tumors and cancer cell lines. The MCF-7 genome has been extensively studied by targeted sequencing analyses of the amplified regions (Hampton et al. 2009; Raphael et al. 2008; Volik et al. 2006) that have revealed complicated sequence structures. By comparing our data with these earlier findings, we found support for all four reported rearrangement points at the *ZNF217* locus (in BAC clone MCF7_1-3F5, (Volik et al. 2006) in the DNA-PET data with the breakpoint coordinates: 1) chr3:-63,998,040/chr20:-52,304,837 (predicted), 2) chr20:+40,267,945/chr20:+ 52,300,747 (predicted), 3) chr20:-40,249,066/chr20:+55,251,653 (predicted), 4) chr20:+55,288,450/chr20:+51,615,194 (PCR validated). To compare our data with 35 rearrangements which have been resolved to the base pair level by Raphael and colleagues (Raphael et al. 2008), we searched for DNA-PET predicted rearrangement points which were within the maximum cPET span of the MCF-7 library (16,217 bp). We found support for 31 out of 35 rearrangement points (Appendix Table 14), of which 15 did not meet our quality criteria (six had a cluster size 2, nine had an anchor span <1000 bp).

Up to now, it is still not clear what events trigger the amplification cascades in cancer genomes. In MCF-7, 26% (268/1047, Table. 3.3) of the rearrangement points were inter-connected into only six highly complex units. The largest of these units involved 205 dPET clusters (Figure. 3.4), including mixed types of mapping patterns, with an overrepresentation of tandem duplications, unpaired inversions, and inter- chromosomal translocations, which were tightly associated with the amplified regions on chromosomes 1, 3, 17, and 20 (Figure 3.23. A). In this complex unit, the SV with the highest dPET cluster count (n=1,176) was mapped to a highly amplified region on 20q13, representing a tandem duplication of a large

fragment (3.67 Mb) at position 51-55 Mb (Figure 3.23. B). Double-probed DNA-FISH experiments validated this rearrangement, and the extensive FISH signals of the mixed probes in linear position and in multiple chromosomal locations indicated that the junction region of this tandem duplication was further multiplied locally as well as dispatched to other chromosomal locations (Figure 3.23. D-F). Thus, this junction appeared as an epicenter for subordinate dPET clusters that were connected to other parts of the genome, either intra- or inter-chromosomally. The dEPT clusters to the left and right of the initial tandem duplication junction were smaller in size than the initial event, but their sum on each side was comparable to the cluster size of the initial event. An inter-chromosomal dPET cluster with 498 dPETs connected chromosome 20 at 55 Mb (left side of the tandem duplication junction in Figure 3.23. C and G) to chromosome 17, while another inter-chromosomal dPET cluster with 370 dPETs connected chromosome 20 at 53 Mb (right side of the tandem duplication junction) to chromosome 17. As the number of dPETs of these two clusters were similar, it is conceivable that the two clusters represent the paired rearrangement points of an inter-chromosomal translocations, which were tightly associated with the amplified regions on chromosomes 1, 3, 17, and 20 (Figure 3.23. A).

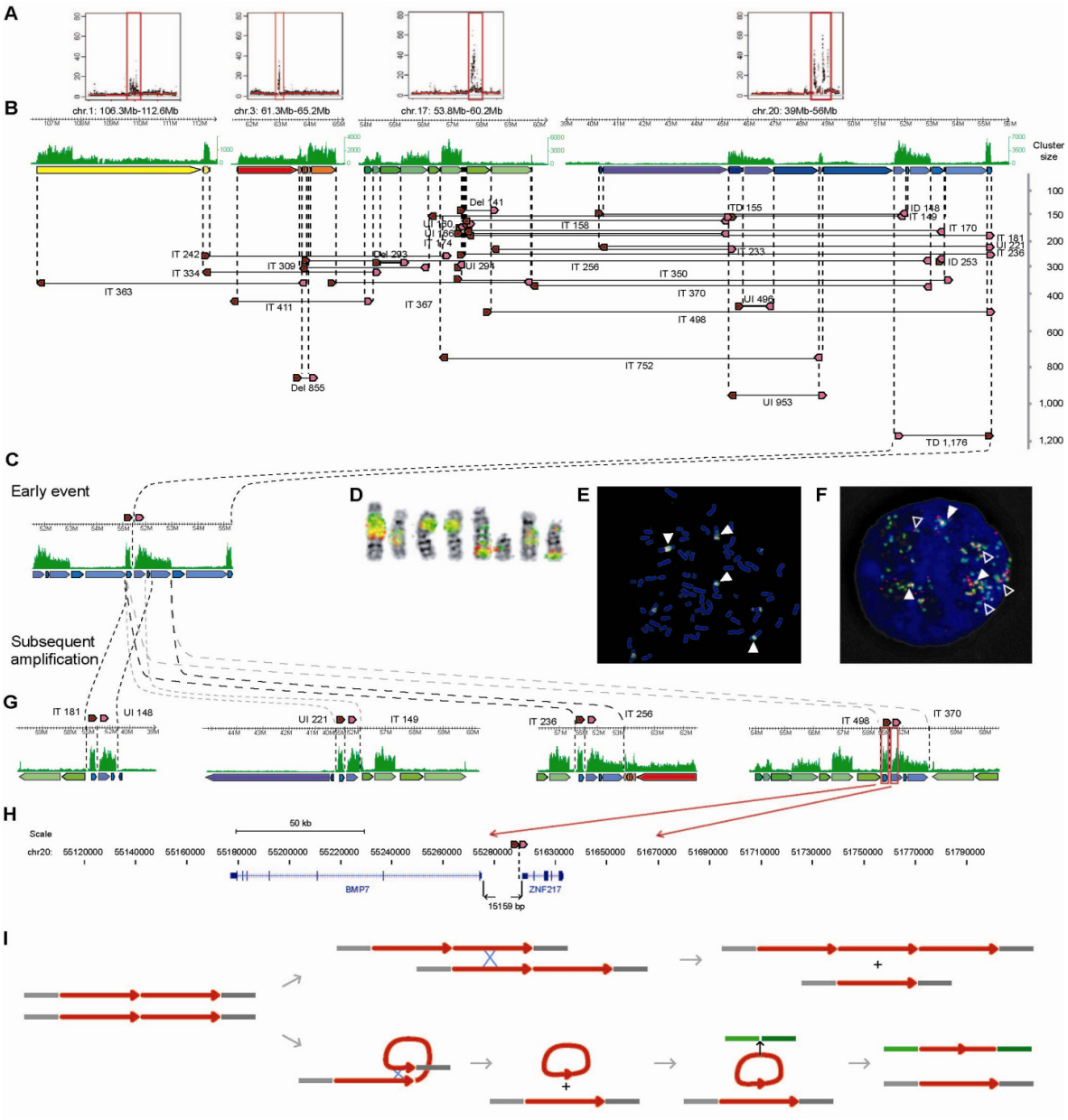

Figure 3.23. Architecture and genealogy of amplifications in MCF-7.

(A) Copy number plots of chromosomes 1, 3, 17, and 20 with amplified regions (red boxes). (B) Concordant tag distributions are shown for amplified genomic regions (top, green track). Genomic segments between predicted breakpoints are indicated by colored arrows (middle) and dPET clusters with cluster sizes greater than 140 are represented by horizontal lines flanked by dark red and pink arrows indicating 5' and 3' anchor regions (bottom). Small to large dPET clusters are arranged from top to bottom. All but three dPET clusters were classified as complex. Mapping characteristics are described by: Del, deletion; IT, isolated translocation; UI, unpaired inversion; TD, tandem duplication. Cluster sizes are given for each cluster. (C) Possible genealogy of amplification. TD1,176 occurred early and subsequent rearrangements have pasted TD1,176 in different genomic contexts (G). (D-F) Double-color FISH using probes flanking TD1,176. Red, chr20:51,920,860-52,096,191; green,

chr20:55,137,293-55,311,637. Double signals (filled arrow heads) indicate the fusion of the two loci and single signals indicate the normal genomic distance (open arrow head). (D) Metaphase chromosomes, (E) metaphase nucleus and (F) interphase nucleus showing amplification and fusion of breakpoint flanking sequences. (H) *BMP7* (left) and *ZNF217* (right) are juxtaposed by the TD 1,176 rearrangement in a distance of 15,159 bp. (I) Models of local and inter-chromosomal amplification. Chromosomes are represented by gray and green horizontal lines. Amplified segment is represented by red arrow. The initial tandem duplication (left) allows local amplification between two sister chromatids or homologous chromosomes (top) or inter-chromosomal translocation (bottom).

In this complex unit, the SV with the highest dPET cluster count (n=1,176) was mapped to a highly amplified region on 20q13, representing a tandem duplication of a large fragment (3.67 Mb) at position 51-55 Mb (Figure 3.23. B). Double-probed DNA-FISH experiments validated this rearrangement, and the extensive FISH signals of the mixed probes in linear position and in multiple chromosomal locations indicated that the junction region of this tandem duplication was further multiplied locally as well as dispatched to other chromosomal locations (Figure 3.23. D-F). Thus, this junction appeared as an epicenter for subordinate dPET clusters that were connected to other parts of the genome, either intra- or inter-chromosomally. The dEPT clusters to the left and right of the initial tandem duplication junction were smaller in size than the initial event, but their sum on each side was comparable to the cluster size of the initial event. An inter-chromosomal dPET cluster with 498 dPETs connected chromosome 20 at 55 Mb (left side of the tandem duplication junction in Figure 3.23. C and G) to chromosome 17, while another inter-chromosomal dPET cluster with 370 dPETs connected chromosome 20 at 53 Mb (right side of the tandem duplication junction) to chromosome 17. As the number of dPETs of these two clusters were similar, it is conceivable that the two clusters represent the paired rearrangement points of an inter-chromosomal translocation that inserted the junction region of the tandem-duplicated segment of chromosome 20 into chromosome 17. Similarly, the same tandem-duplicated junction appeared to be, by a separate translocation, connected to chromosome 17 and 3

(cluster sizes 236 and 256, respectively). Further shorter segments of this tandem duplication junction were disseminated to other locations on chromosomes 17 and 20 (by two pairs of dPET clusters with 221/149 and 181/148 dPETs, respectively; Figure 3.23.G). It is worth noting that the sum of relative copy numbers of the rearranged segments inferred by these four subordinate dPET clusters coincides well with the copy number of the tandem duplication on chromosome 20 (see below), suggesting that this tandem duplication junction is the origin for the subsequent amplification and dissemination. The locus on chromosome 20 gains more complexity by long distance unpaired inversions (45.2-48.7 Mb and 45.8-47 Mb) of high cluster size and further connections to the region on chromosome 17 (56-60 Mb). This suggests that the tandem duplication junction was the origin for the subsequent amplification and dissemination.

The dPET connectivity and PET counts together delineate a possible genealogy of rearrangements in the MCF-7 genome. We hypothesize that this 3.67 Mb segment in tandem duplication was the first rearrangement; it then probably created a state of genomic instability, and triggered a cascade of subsequent rearrangements that centered around the junction point of the initial tandem duplication, probably by providing the substance for NAHR. Such recombination could take place between sister chromatids to result in further linear amplification of this duplicated segment, or intrachromatid generating potential "double minute" constructs that could be further amplified and eventually inserted in other parts of the genome (Figure 3.23. I).

The extensive proliferation of this rearranged structure suggests that some driver element(s) were created to favor the selection of this junction segment during the evolutionary course of this genome. This tandem duplication juxtaposes the *BMP7* gene immediately downstream of the *ZNF217* gene (Figure 3.23. H), and this two-gene construct is intact in the minimal core segment that has been amplified most extensively, suggesting that

it was advantageous for its extensive amplification. Quantitative reverse transcription PCR (qRT-PCR) of *BMP7* and *ZNF217* proved that both genes are highly expressed in MCF-7 (Figure 3.24), suggesting that MCF-7 cells achieve high expression of these two genes through high gene copy numbers. It is still not entirely understood how this two-gene locus could functionally achieve such superior propagation in MCF-7.

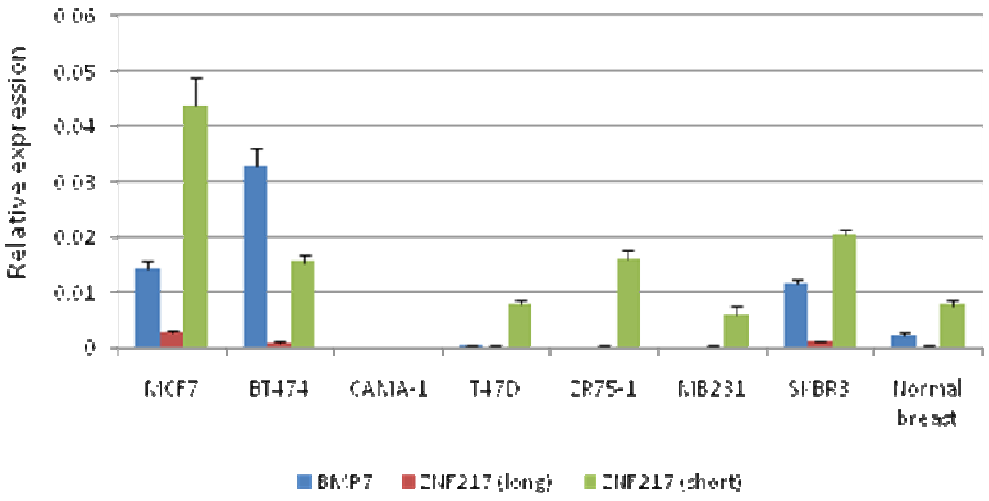

Figure 3.24. RT-PCR of *BMP7* and *ZNF217* in breast cancer cell lines and normal breast.

Expression relative to beta-actin (y-axis) of three independent experiments is shown. *ZNF217* (long) corresponds to UCSC known gene uc010gij.1 and *ZNF217* (short) corresponds to NCBI RefGene NM_006526.2.

Genomic architecture of amplified regions in SKBR3

Similarly, the SKBR3 genome also has a few complex units of rearrangements located in highly amplified regions. The largest complex unit consists of 50 rearrangements (Fig. 3.6) The rearrangement with the highest dPET cluster count (n = 624) in this genome is also a large tandem duplication (9.1 Mb) and mapped to chromosome 8 at location 72.8-82 Mb (Figure 3.24). It is also involved in highly amplified regions and is connected to other rearrangement sites. Based on dPET connectivity and PET counts, we reconstructed the amplified regions that involved at least two levels of subordinate tandem duplications and an inter-chromosomal translocation that connected chromosome 8 (71.4-82 Mb, 87.2-92.6 Mb, 109.8-129.2 Mb) to chromosome 17 (34.5-37.8 Mb) (Figure 3.25). Similar to the amplicon regions in MCF-7, the SKBR3 data implies that the fusion point created by the tandem duplication occurred early in the genealogy of this breast cancer genome and that subsequent events have led to the amplification of that fusion junction.

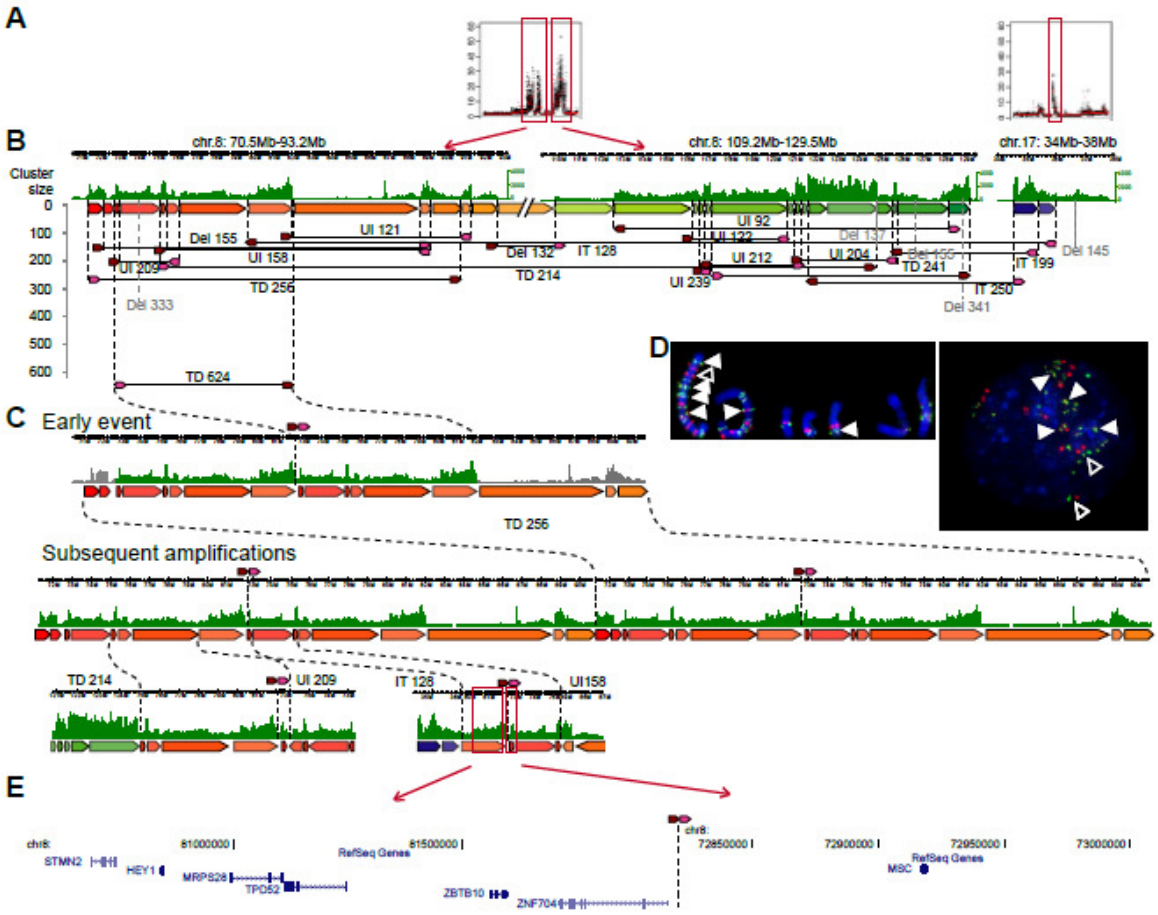

Figure 3.25. Architecture and genealogy of amplifications in SKBR3.

(A) Copy number plots of chromosomes 8 (left) and 17 (right) with amplified regions (red boxes). (B) Concordant tag distributions (in green) are shown for amplified genomic regions on chromosomes 8 and 17 (top). Genomic segments between predicted rearrangement points are indicated by colored arrows (middle) and dPET clusters with cluster sizes greater than 90 are represented by horizontal lines flanked by dark red and pink arrows (bottom). Abbreviations are as described in Fig. 4. Deletions <50 Kb are in gray and were not considered for genomic segmentation. (C) Possible genealogy of amplification. TD624 occurred early (top left, rearrangement point is represented by dark red and pink arrows). Subsequent rearrangements have pasted TD624 in different genomic contexts and thereby amplified the rearrangement point. (D) Double-color FISH experiments using probes flanking TD624. Red represents chr8:72,695,393- 72,880,900 and green chr8:81,904,336-82,095,683. Note the repetitive linear sequence of the two loci with double signals (filled arrow heads) indicating the fusion of the two loci and single signals indicating the normal genomic distance (open arrow head). (E) Genes flanking TD624 with *STMN2*, *HEY1*, *MRPS28*, *TPD52*, *ZBTB10*, and *ZNF704* on the left and *MSC* on the right.

Genomic architecture of amplified regions in primary tumor

Primary tumor genomes also have extensive amplifications (Figure 3.6). For instance, breast tumor 14 displayed a local amplification on chromosome 9p (Figure 3.26. A) where 8 dPET clusters (PET counts >8) were connected to this amplified locus, including four large tandem duplications, two unpaired inversions, and two deletions (Figure 3.26. B). The deletion with the largest cluster count excises exons 2 to 6 of *KDM4C* (Figure 3.26. C) and the exon 1-7 fusion was validated by RT-PCR (Figure 3.26. D). *KDM4C* (also known as *GASC1*) has been described as an oncogene in breast cancer (Liu et al. 2009). If translated, this truncated protein would lack the entire JmjN domain; have a partial JmjC domain, and an intact PHD-finger for possible new function.

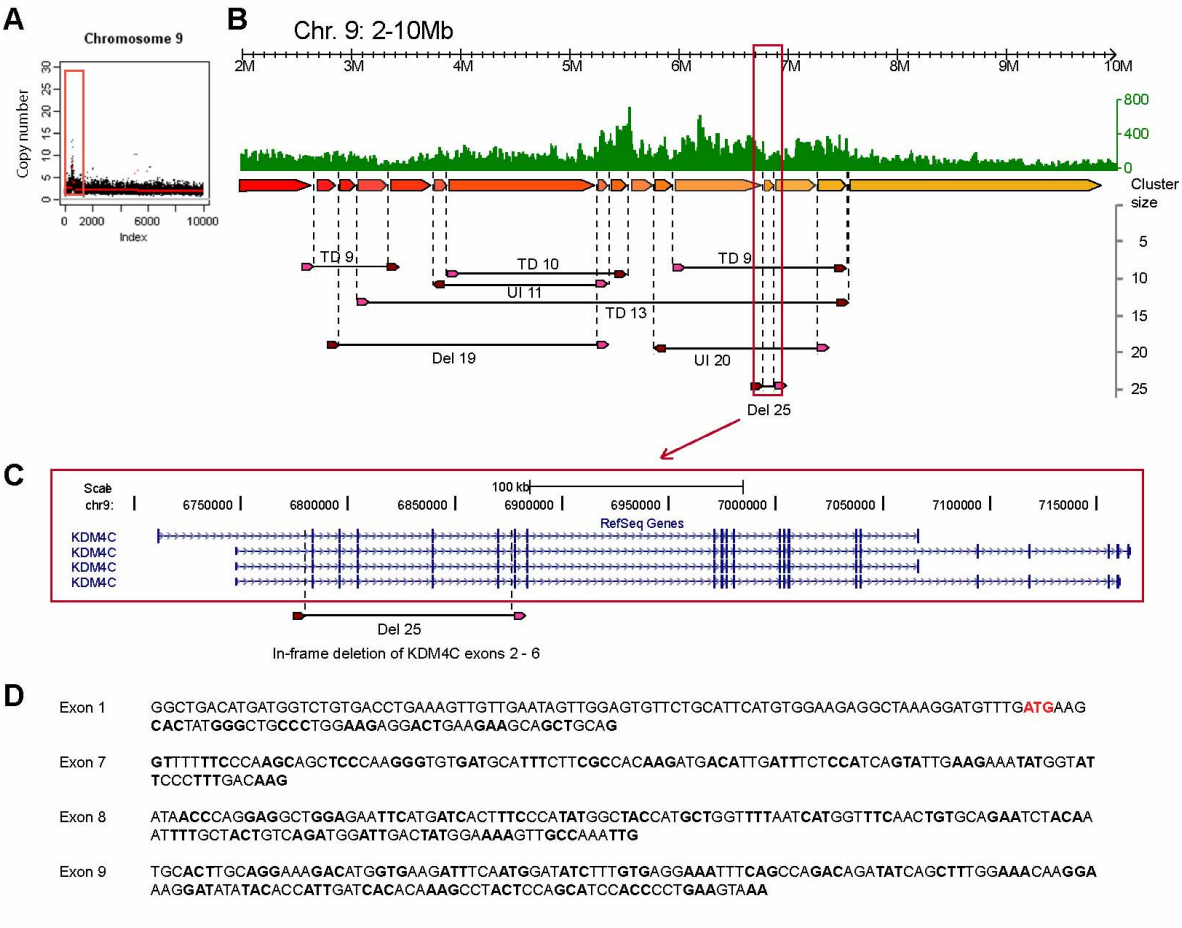

Figure 3.26. The architecture of an amplified region in primary breast tumor 14.

(A) Concordant tag based copy number estimate for chromosome 9 indicates an amplification of the distal region of 9p. (B) Concordant tag distribution of chromosome 9 position 2-10 Mb (top, green track). Genomic segments between predicted breakpoints are indicated by colored arrows (middle) and dPET clusters with cluster sizes greater than eight are represented by horizontal lines flanked by dark red and pink arrows (bottom). Abbreviations for mapping characteristics of dPET clusters are described in Figure 6. (C) Genomic structure of *KDM4C*. Location of amplified deletion (Del25) is indicated by dashed vertical lines. (D) Sequencing result of RT-PCR confirms the in-frame deletion transcript with the more upstream located exon 1.

Breakage-fusion-bridge cycle in cancer genome

We observed many long distance unpaired inversions in the breast and gastric cancer genomes, which could indicate the inversion of whole chromosomal arms, large inversions or inverted insertions are involved in further rearrangements, or a failure to detect the paired rearrangement point that would classify the event as an inversion. On the other hand, unpaired inversions with a relatively short distance between their breakpoints could occur when a DNA double strand break results in a truncated chromosome, followed by the replication of the DNA and the joining of the two neighboring ends by a DNA repair mechanism in a head to head or tail to tail fashion (resulting in a fusion of + and – strand of the sister chromatids, Figure 3.27. A). Due to the fusion, the two sister chromatids cannot be separated in mitosis and a new break could occur to initiate a new fusion. This mechanism has been described as breakage-fusion-bridge (BFB) cycles which was originally proposed by Barbara McClintock in 1941(reviewed in (Tanaka et al. 2009). In this model, a double-strand chromosome break generating a free DNA end is followed by DNA synthesis and sister chromatid formation, resulting in two identical free DNA ends. The pair of sister chromatids then fuse to each other in order to eliminate the free ends, which otherwise may trigger cell death. Following chromatid separation during mitosis, an anaphase bridge is formed, resulting in a further doublestrand break and re-initiation of the breakage-fusion-bridge cycle. A distance of a few kilobases between head to head fusion points has been reported (Bignell et al. 2010; Lo et al. 2002; Okuno et al. 2004). Gastric tumor 17, which had the most rearranged and amplified genome among the four gastric tumor samples (Figure 3.6) showed an accumulation of short distance unpaired inversions in the amplified regions on chromosomes 5, 11, 12, and 18 (Figure 3.27. B). This pattern can be explained by BFB cycles which are known to result in amplifications (Tanaka et al. 2009) (Figure 3.27. C-D). The dPET counts implied that a translocation between chromosomes 5 and 18 (cluster count,

n=382) preceded a double strand break and a subsequent tail to tail fusion of chromosome 5 at 39.2 Mb by an unpaired inversion (cluster count, n=118; Figure 3.27.). Further breaks and fusions amplified the chromosome 5 and 18 segments. A break in a postulated second BFB cycle resulted in two sister chromatid fusions, which showed a larger distance between their breakpoints of 390 and 450 kb, respectively, and involved a loss of 1.5 Mb. The data imply the propagation of different populations of rearranged chromosomes, which together result in the amplification of the two loci. We observed a larger number of small (<10 kb) unpaired inversions per chromosome in the gastric cancer samples than in the breast cancer samples (p=0.00587). This might indicate that BFB cycles are more characteristic for gastric rather than for breast cancer.

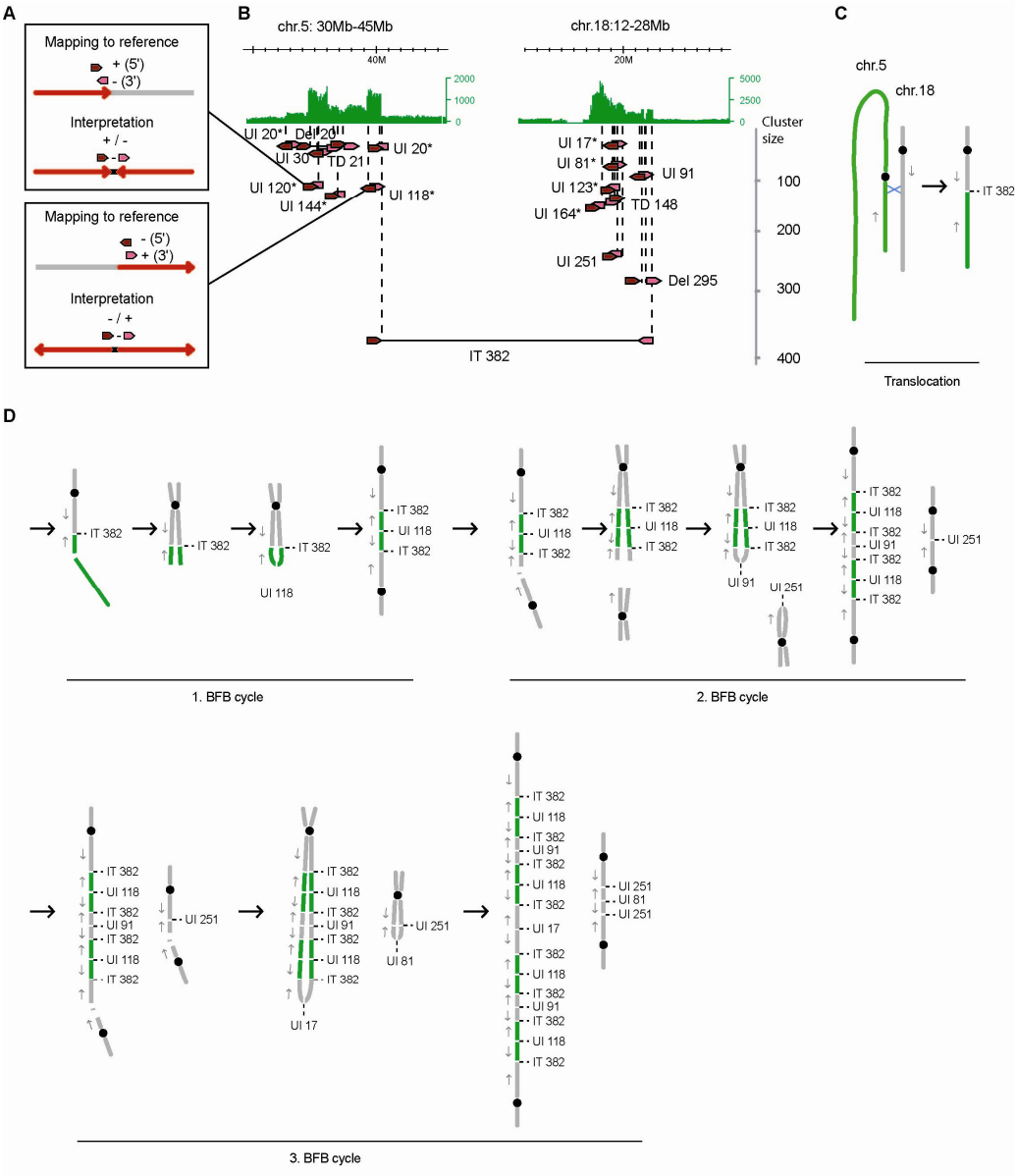

Figure 3.27. Accumulation of short span unpaired inversions in amplified regions of gastric tumor 17.

(A) PET mapping pattern of short span unpaired inversions and the interpretation. The mapping of a 5' anchor (dark red arrow) to the + strand and a 3' anchor (pink arrow) to the – strand indicates a head to head fusion (red arrows) with increasing chromosomal coordinates closer to the breakpoint (top) and a 5' – strand/3' + strand mapping indicates a tail to tail fusion with decreasing chromosomal coordinates closer to the breakpoint (bottom). UI120 and UI118 in (B) are examples for head to head and tail to tail fusions, respectively. (B) Amplifications on chromosomes 5 and 18 of gastric tumor 17 are indicated by concordant tag counts (green). Cancer structural rearrangements with dPET cluster sizes >15 are indicated by dark red and pink arrows for 5' and 3' anchors, respectively. Abbreviations and figure structure are described in legend of Figure 6B. Unpaired inversions with a breakpoint distance <40 Kb are indicated by asterisks. (C) Schematic representation of an isolated translocation between chromosome 5 (green) and 18 (gray). Black circles represent centromeres, blue X represents site of recombination. Gray arrows indicate the direction of increasing genomic coordinates. (D) Interpretation of accumulated short unpaired inversions in amplifications by BFB cycles.

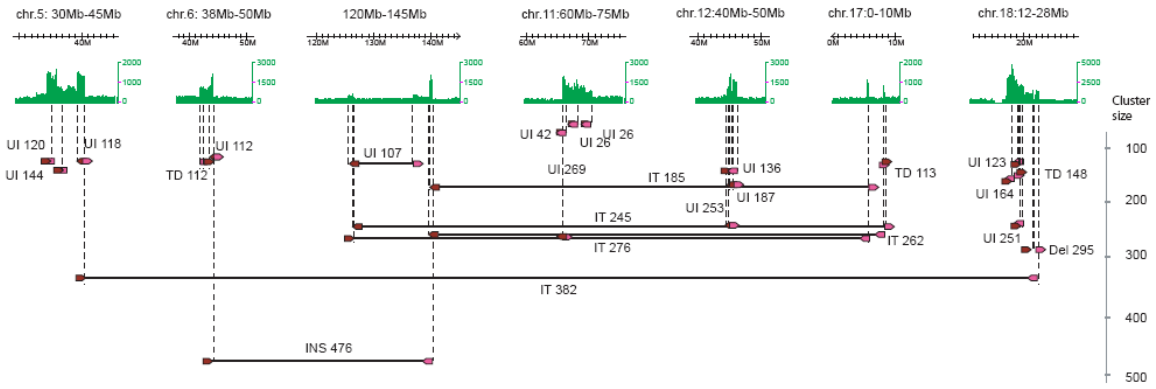

Figure 3.28. Architecture of amplifications of gastric tumor 17.

Concordant tag distributions of amplified regions on chromosomes 5, 6, 11, 12, 17, and 18 are represented by green tracks. Rearrangement points are represented by dark red and pink arrows connected by horizontal lines. Rearrangement points of cancer specific dPET clusters with a size >100 are arranged according to dPET cluster size from top (small) to bottom (large). Three unpaired inversions on chromosome 11 with cluster sizes <100 have been included to explain copy number changes. Deletions <20 kb are not displayed.

Discussion

We have comprehensively characterized SVs of 15 human cancer genomes and 2 normal human genomes by paired-end-tag sequencing and mapping analysis. The use of a 10 kb insert size for DNA-PET analysis allows the identification of breakpoints within repetitive or homology containing regions of a few kilobases in size and results in a higher physical coverage compared to small insert libraries with the same sequencing effort. The recently reported study by Stephens et al. (Stephens et al. 2009) on breast cancer genome structures using short DNA fragments may have insufficient physical coverage of the genome to map rearrangements in complicated genomic regions without dramatically increasing the sequencing coverage. Thus, long span DNA paired-end approaches such as outlined here represent a parsimonious and cost effective approach to comprehensively map structural mutations in cancers.

All the five primary breast tumors are belonging to basal-like breast cancer. The characterization of basal-like breast cancer is the absence of oestrogen receptor (ER) expression, the lack of *ERBB2* gene amplification, and a high mitotic index. The rapid fatal clinical course is the consequent of no approved targeted therapy options and poor response to standard chemotherapy. The poor understanding the genetic events of this tumor subtype limited the clinical progress (Carey et al. 2006). Recently, four DNA samples from a African-American patient with basal-like breast cancer: peripheral blood, the primary tumor, a brain metastasis and a xenograft derived from the primary tumor were sequenced by next generation sequencing technologies (454 and Illumina, (Ding et al. 2010). The comprehensive analysis of these four samples identified 50 novel somatic point mutations and small indels in coding sequences, RNA genes and splice sites as well as 28 large deletions, 6 inversions and 7 translocations. Two de novo mutations, one was a missense mutation (T708I) in *SNED1*, with a mutant frequency of 37%, the other was a silent mutation (N2483) in *FLNC* with a mutation allele frequency of 18% were found only in the metastatic tumor, but not in primary or xenograft tumor genomes. In addition, a 26 kb deletion in *MECT* was only identified and validated in the metastasis, suggestion its de novo nature in this sample. The results also showed that xenograft retained all primary tumor mutations and displayed a mutation enrichment pattern that resembled the metastasis (Ding et al. 2010). Another notable finding is that two overlapping large deletions (538 kb and 515 kb in length) on chromosome 5, affecting *CTNNA1* along with *LRRTM2*, *MATR3*, *SNORA74A* and *SIL1*, were found to be present in all three tumor samples. *CTNNA1* was shown to be very important for the cell adhesion in breast cancer cells (Bajpai et al. 2009) and increased *in vitro* tumorigenic characteristics (Plumb et al. 2009). The bi-allelic deletion might have important function. This study demonstrated that, although additional copy number changes, structural variations and somatic mutations do occur during the progress of the disease, most

of the original mutations and structural variations present in the primary tumor are propagated.

From the data generated in this study, some characteristic patterns of SVs in cancer genomes including primary tumors and cancer cell lines emerged. Inversions, insertions, and deletions are more commonly seen in germ line SVs; whereas somatic rearrangements present in cancer genomes are overrepresented in tandem duplications, unpaired inversions, isolated translocations, and in amplified complex regions. Such distinction is likely due to mechanistic differences: SVs with germ line origins are meiotic recombinants, whereas somatic SVs may use a variety of mechanisms including mitotic DNA repair, transcription-mediated recombination, and generation of double-minute structures (Gu et al. 2008; Kuttler et al. 2007; Lin et al. 2009; Murnane et al. 2004).

The precise and quantitative connectivity assessment of fusion points in amplified regions by dPET clusters provided an opportunity to delineate the genealogy of amplifications in cancer genomes. In the samples of breast and gastric cancer that we examined, we have gathered evidence to show that large tandem duplications as well as unpaired inversions appear to be early events triggering a subsequent cascade of extensive amplification centered around the junction region. Though it remains a possibility that the tandem duplication may simply function as a "marker" for regional genomic instability remains a possibility, the propagation of the precise tandem duplication through progressive amplification in several cancer genomes suggests that particular tandem duplications have "driver" function. The evolutionary signature of an initiator-amplification cycle is further supported by the observation in the K562 cell line where the *BCR-ABL1* balanced translocation between chromosomes 9 and 20 has been subsequently amplified in other parts of the genome (Figure 2.15).

It is not clear what mechanisms dictate the initial structure of the early events for amplification and why epithelial cancer genomes favor local duplication whereas leukemia genomes prefer inter-chromosomal translocation. The difference at this level may be associated with the spatial state of chromosome conformation in particular cell types and the microenvironment of selective pressure where the primary cells reside.

Complex rearrangements and extensive amplification are much more abundant in the cancer cell lines than in primary tumors. This is likely due to additional rearrangements that are acquired during *in vitro* passages of the cell lines. However, detailed analysis of these rearrangements could provide an evolutionary model that "amplifies" the functional importance of any particular rearrangement. To a degree, such a rearrangement map of a highly passaged cell line may represent a steady state of genome fitness for a specific cancer, especially for *in vitro* conditions.

It is well known that cancers from different lineages such as epithelial and mesenchymal origins harbor very different genetic rearrangements: balanced translocations are predominantly found in mesenchymal cancers, whereas complex rearrangements are a hallmark of mature epithelial cancers. Our focus on two epithelial cancers, breast and gastric, was an attempt to assess the differences between two distinct epithelial cancers that arise from very different epidemiologic etiologies. We found, within the limitations of sample size, that breast and gastric cancer genomes have some comparable structural characteristics. Both cancers show an enrichment of tandem duplications, unpaired inversions, isolated translocations and complex rearrangements. In breast and gastric cancer, tandem duplications are larger than other SV categories and have a higher chance of enclosing genes. However, gastric cancer rather than breast cancer shows signatures that are compatible with the breakage-fusion-bridge (BFB) model, which might suggest different mechanisms of genome instability.

The mapping of primary tumor genomes demonstrated in this study validated the feasibility of using large span paired-end-tag sequencing to characterize clinical samples for genome structural variations. With further optimization for the current prototype of DNA-PET analysis and continuous drop of sequencing cost, we expect this approach to be sufficiently robust and cost effective to be applied in clinical settings for genetic diagnostics of cancer patients and other genetic disease patients.

Chapter Four: Conclusions

Summary

In this thesis, I have demonstrated that long span DNA-PET (10 kb) is a powerful tool to study human cancer genomes. The use of a 10 kb insert size allows the identification of breakpoints within repetitive or homology-containing regions of a few kilobases in size and results in a higher physical coverage compared with small insert libraries (1 kb) with the same sequencing effort. The short span DNA-PET had only advantages to identify deletions smaller than 5 kb. In addition, our data also demonstrated that 10 kb library had a comparable resolution (114 bp vs. 377 bp in 1 kb and 10 kb library) in predicting breakpoint locations to a distance that can be amplified by PCR. The 20 kb insert size library had a slight advantage in discovering inversions and unpaired inversions compared to 10 kb insert size library but displayed a low sensitivity in identifying small SVs of various categories. Moreover, the construction of libraries with 20 kb inserts requires more genomic DNA starting material. The detailed characterization of SVs by long span DNA-PET libraries showed many new sub-types of insertions, which could help in understanding the effect and genesis of insertions in human cancer genomes.

We have applied this long span DNA-PET approach to comprehensively characterize the SVs of two epithelial cancers, breast and gastric cancer. Fifteen cancer genomes and two normal genomes were analyzed and we used a filtering approach to strongly enrich for somatic SVs in the cancer genomes. Our analyses revealed that most inversions, deletions, and insertions are germ-line SVs, whereas tandem duplications, unpaired inversions, interchromosomal translocations, and complex rearragements are over-represented among somatic rearrangements in cancer genomes. We demonstrate that the quantitative and connective nature of DNA-PET data is precise in delineating the genealogy of complex rearrangement events, we observe signatures that are compatible with breakage-fusion-bridge

cycles, and we discover that large duplications are among the initial rearrangements that trigger genome instability for extensive amplification in epithelial cancers.

With the rapid development of next generation sequencing technologies, whole genome sequencing has become an invaluable tool for obtaining a complete understanding of human genomic variation. In the future, personal genomic information will gain importance to tailor an individual's medical care. With further optimization for the current prototype of DNA-PET analysis, we expect this approach to be sufficiently robust and cost effective to be applied in clinical setting for genetic diagnostics of cancer patients and other genetic disease patients.

Further development of NGS platforms

There are three major limitations in current second generation sequencing technologies: (1) the read length is short; at present, NGS only can provides 50-500 continuous basepair reads, much shorter compared to traditional Sanger sequencing (1000-1200 bp); (2) the coverage which is defined as the number of short reads that overlap each other within a specific genomic region is uneaven; since NGS technologies only can produce short reads and repetitive sequences affect the mappability of short sequences, coverage becomes a very important issue, especially for accurate assembly of the genomic sequence; (3) mappable, short reads generated by NGS technologies create many sequences that cannot be interpreted or "mapped" unambiguosly to the reference DNA or be accurately assembled. This is because a number of short reads match with many different genomic regions and are therefore not unique to any specific genomic region.

The second generation sequencing protocols use an amplification step of DNA fragments by emulsion or cluster PCR, to make the light signal strong enough for reliable base detection by the CCD cameras. Although the PCR amplification has revolutionized

DNA analysis, in some cases, it may introduce error base sequences or favor certain sequences over others, thus changing the relative frequency and abundance of various DNA fragments that existed before amplification (Pareek et al. 2011a). Thus, the sequence determined directly from a single DNA molecule without PCR amplification will overcome this problem. The sequencing of single DNA molecules is now called as the "*third generation sequencing technology*" (Schadt et al. 2010). Heliscope^{TM} single molecule sequencer is one of the first techniques for sequencing from a single DNA molecule and its principle relies on "*true single molecule sequencing*" (tSMS) technology. In this tSMS technology, a poli-$_{(A)}$ tail is introduced into DNA during library preparation so that the DNA molecules hybridize to the poli-$_{(T)}$ oligonucleotides which are attached to a flow cell and simultaneously get sequenced in parallel reactions. The sequencing cycle consist of DNA extension with one fluorescently labeled nucleotide, out of four, followed by nucleotide detection with the Heliscope sequencer. The subsequent chemical cleavage of fluorophores allows the next cycle of DNA elongation to begin with another fluorescently labeled nucleotide, which enables the determination of the DNA sequence (Pareek et al. 2011b). The Heliscope sequencer can sequence up to 28 Gb in a single sequencing run (8 days) with a maximal length of 55 bases.

Single DNA molecule sequencing technology can read through DNA templates in real time without amplification and provides accurate sequencing data with potentially long-reads, therefore, several single-molecular DNA sequencing technologies are currently under development (Zhang et al. 2011). Several single molecule sequencing emerged over the last years: (1) Flourescence-based single-molecule sequencing is performed by identifying nucleotides which are phospholinked with distinctive colors. During the synthesis process, fluorescence emitted as the phosphate chain is cleaved and the nucleotide is incorporated by a polymerase into a single DNA strand. (2) Nano-technologies for single-molecule sequencing are based on the principle that thousands of nano-tunnels on a chip can be used to monitor the

movement of a polymerase molecule on a single DNA strand during replication. This principle comes from the observation that when a DNA strand is pulled through a nanopore by an electrical current, each nucleotide base (A, T, C, G) creates a unique pattern in the electrical current. This unique nanopore electrical current fingerprint can be used for naopore sequencing. Electrical base detection reads through the stretched and immobilized strand of DNA molecules on conductive surfaces by multiple nano-knife edge probes. Each non-knife edge probe specially recognizes only one nucleotide for single-molecule sequencing. (3) Other developing approaches for single-molecule sequencing include electron microscopy, ion sensor or sequencing-by-hybridization technology based on known reference sequences.

In summary, the third generation sequencing technologies may offer the following advantages over second generation sequencing: i) no PCR amplification bias, ii) higher throughput, iii) faster turnaround time (e.g., sequencing metazoan genomes at high fold coverage in minutes), iv) longer read length to enhance de novo assembly and enable direct detection of haplotypes and even whole chromosome phasing, v) higher consensus accuracy to enable rare variant detection, vi) small amounts of starting material (theoretically only a single molecule may be required for sequencing) and vii) low cost.

Besides the sequencing technologies, the full benefits of NGS will not be achieved until extremely high-performance computing and intensive bioinformatics support is available. Typically, tens or hundreds of Gbp short reads can be generated during each run in any given NGS platform. Given the vast amount of data produced by NGS, developing a massive data storage and management solution and creating informatic tools to effectively analyze data will be essential to the successful application of NGS technology (Zhang et al. 2011).

Challenges in cancer genome sequencing

The quantity, quality and purity of cancer samples are normally quite different from the peripheral blood samples which are commonly used as germline control. Solid tumors are complex mixtures of cells including non-cancerous fibroblasts, endothelial cells, lymphocytes and macrophages that often contribute more than 50% of the total DNA. This mixture can mask the signal from the cancer cells and complicate the inter- and intra-tumor comparisons (Meyerson et al.). In addition, solid tumors are often highly heterogeneous and composed of different clones that have different genomes. The detection of somatic mutations in cancer needs mutation calling in both the tumor and the matched normal DNA, coupled with the comparison to a reference genome. There are two kinds of false positive genome variation calls: inaccurate detection of an event in the tumor, when the tumor and normal are both wild-type; detection of a germline event in the tumor but failure to identify it in the normal genome. The first type of error can be induced by sequencing error, incorrect local alignment or discordant alignment of paired reads. The second type of false positive mutation calls mainly come from insufficient coverage to fail detect the germline alleles that differ from the reference sequence (Meyerson et al. 2010).

Single-cell genomic methods have the capacity to resolve complex mixtures of cells in tumors. Recently a new technology called single nucleus sequencing (SNS) had been developed by Navin and colleagues (Navin et al. 2011b), in which three technologies are combined including next generation sequencing (NGS), fluorescence-activated cell sorting (FACS) and whole genomic amplification (WGA). The SNS method has been used to investigate tumor population structure and evolution in two human breast cancer cases. Three distinct clonal subpopulations isolated from 100 single cells of a polygenomic tumor may represent sequential clonal expansions. One hundred single cells from a monogenomic primary tumor and its liver metastasis showed a single clonal expansion formed by the

primary tumor which seeded the metastasis. This study demonstrated that we can make inferences about the evolution and spread of cancer by the copy number profiles from sequencing a single cell. The identification of pseudodiploid cells showed that single cell sequencing methods can identify cell types which could not be detected by previous methods (Navin et al. 2011b).

Single cell sequencing methods provide an unprecedented view of the genomic diversity within tumors and provide the means to detect and analyze the genomes of rare cancer cells. Although the cancer genome studies with bulk tissue samples can provide a global spectrum of mutations, they cannot determine whether all of the tumor cells contain the full set of mutations, or different subpopulations contain subsets of these mutations that in combination drive tumor progression. Furthermore, single cell sequencing probably can greatly improve our fundamental understanding of how tumor evolve and metastasize. The future medical application of single cell sequencing will be in early detection, monitoring circulating tumor cells (CTCs) during treatment of metastatic patients and measuring the genomic diversity of solid tumors. However, there are still many challenges ahead of single cell sequencing: i) the low coverage of the human genome (6%); ii) the detection of copy number variation only, but not structural variations due to chimeric products generated by amplication; iii) the need to profile hundreds of single cells quickly and at a reasonable cost; iv) the limitation on frozen tumor samples but not paraffin embedded samples. When future innovations allow whole genome sequencing of single tumor cells, oncologists will also be able to obtain the full spectrum of genomic sequence mutations in cancer genes from scarce clinical sample. These methods are likely to improve all three major themes of oncology: prognostics, diagnostics and chemotherapy, ultimately improving the treatment and survival of cancer patients (Navin et al. 2011a).

In conclusion, all the pioneer studies indicated that cancers typically carry a few consistent and functionally characterized abnormalities, companied with tens to thousands of other changes that are rear or unique to the individual tumor with little known. We should understand which genes contribute to tumorigenesis or progression and how these changes happened during tumor evolution and interact with the tumor microenvironment ─ and therefore how it regulates each tumor's behavior and response to therapy. All these understanding will help to develop the therapeutic approaches targeting specific cellular pathways.

Chapter Five: Materials and methods

Note: Except for the very first few libraries run by Forster City, ABI, USA, all sequencing described here was performed by the Sequencing Team of Genome Technology and Biology by Wei Chia-Lin and Xiaoan Ruan, Genome Institute of Singapore, Singapore. The members of the sequencing team are Herve Thoreau (lab manager), Dawn Choi, Low Hwee Meng, Ong Chin Thing (Jo), Leong See Ting, Adeline Chew, Lee Yen Ling, Poh Tong Shing and Lim Kian Chew.

Materials and Methods used in chapter 2

Cell culture and genomic DNA extraction

MCF-7 (ATCC# HTB-22^{TM}), HCT116 (ATCC# CCL-247^{TM}) and K562 (ATCC# CCL-243^{TM}) were grown under standard culture conditions and harvested at log phase. The genomic DNA was extracted by Blood & Cell Culture DNA Midi Kits (Qiagen) according to the manufacturer's instruction.

Library construction and sequencing

We randomly sheared up to 50 µg of genomic DNA to 1 kb, 10 kb and 20 kb fragments by HydroShear (Genomic Solutions Inc) according to the manufacturer's instruction. The fragmented DNA was methylated using *EcoP15I* (NEB) and end polished by End-It^{TM} DNA End-Repair kit (Epicentre Biotechnologies). SOLiD *EcoP15I* CAP adaptor (Applied Biosystems) which contains the *EcoP15I* restriction site was blunt-end ligated to the two ends of DNA fragments and the ligation products were size-selected on an agarose gel. The small DNA fragments (1 kb) were purified by QIAquick Gel Extraction Kit (Qiagen) and large DNA fragments (10 kb and 20 kb) were purified by QIAEX II Gel Extraction Kit (Qiagen). Up to 1 µg of gel selected DNA fragments were circularized with the biotinylated SOLiD Internal Adaptor (Applied Biosystems) at 0.1 ng/µl final DNA concentration and the uncircularized DNA fragments were removed by Plasmid-Safe^{TM} ATP-Dependent DNase (Epicentre Biotechnologies). The remaining circularized DNA fragments were digested by

EcoP15I (NEB) to release the 25-27 bp di-tags from genomic DNA fragments. Di-tag constructs were end repaired, bound to streptavidin beads and washed. SOLiD sequencing adaptors (Applied Biosystems) were ligated and di-tag constructs were amplified with SOLiD PCR primers (Applied Biosystems) by a 16-cycle PCR. High-throughput sequencing of the 2 x 25 bp libraries was performed on SOLiD sequencers according to the manufacturer's recommendations (Applied Biosystems). ***Note: libraries IHH026, IHK016 and IHK017 were constructed by Zhang Zhenshui; IHK002, IHK004 and IHK007 were constructed by Poh Hui Mei, respectively. Genome Institute of Singapore, Singapore.***

PET sequencing analysis

Mapping, pairing and rescuing of sequence di-tags

The paired tags designated as R3 and F3 were mapped individually to the reference sequence (hg18, NCBI build 36) in color space by the ABI SOLiD pipeline Corona Lite (Applied Biosystems). Contigs of the reference sequence with unresolved location (random_chr) and alternative MHC haplotypes were excluded from the reference for mapping since they caused ambiguous mapping due to high sequence similarity to other sequences in the reference. The pairing and rescuing procedure can be divided in four steps. 1) For each bead (di-tag amplicon unit on a sequencing slide) collect and match all R3 and F3 mapping locations with up to 2 mismatches (for a 2x25 bp library). 2) If both R3 and F3 tags are present and at least one has hit(s) to the reference go to pairing, otherwise discard the bead. 3) Pairing: if both tags have matches to the reference, try all R3/F3 combinations to see if there is a single pair combination with correct orientation and specified insert range. If there is such a combination report it as AAA indicating that both tags are on the same chromosome, same strand, read in the same direction, are in the correct order [R3 is 5' of F3]; and within 20 kb of each other (for 10 kb library, 5 kb for 1 kb library and 40 kb for 20 kb library); if there is more than one

pair of AAA, discard the bead. If there is no AAA, go to rescue. 4) Rescue: anchor each hit to the reference and search nearby sequence for AAA hits with up to 4 mismatches. If a single AAA can be found, report it as AAA; if there are more than one pair of AAA, discard it. If no AAA can be found, paired reads are classified according to ABI SOLiD nomenclature. Redundant PETs which had the same starting points for both tags were believed to be derived from the same PCR product of the library amplification step and were removed from further analysis. Non-AAA PETs were defined as redundant if the starting points of both tags were within +/- 2 bp.

PET classification

Based on ABI SOLiD pairing report, we further separated all the PETs into concordant PETs (cPETs) and discordant PETs (dPETs). cPETs were defined as those where both tags mapped to same chromosome, same strand, in the correct 5' to 3' ordering and within expected span range. The span range was determined by the following process. The span distribution of all AAA PETs was plotted. The distribution was smoothened by averaging values across overlapping windows with the size of 0.1 x standard deviation (μ) of the average span (within 0-20,000 bp, where μ above and below average was calculated separately). Smoothening for each point x was done by averaging the values across the window $[x - \mu/2 , x + \mu/2]$. Then the gradient at each position along the distribution was calculated by gradient(x) = (*smoothened_dist* $(x + \mu/2)$ - *smoothened_dist* $(x - \mu/2)$) / μ. The minimum span point should be the point left of maximum gradient where gradient reaches 0. The PETs which were rejected by cPET criteria were classified as dPETs. These were further split into five distinct categories; (i) two tags mapped on different chromosomes, (ii) two tags mapped on the same chromosome, but different strand, (iii) two tags mapped on the same chromosome, but wrong order (5' downstream of 3'), (iv) two tags mapped on the same chromosome, same strand,

correct order, but with larger span distance than 1.1x the maximum library size, (v) two tags mapped on same chromosome, same strand, correct order, but with smaller span distance than the minimum library size.

Clustering of dPETs

Discordant PETs may result from either inherent SVs in the sequenced genome or due to ligation errors in the library construction process. To filter out those from ligation errors, we identified those discordant PETs that occurred together to form clusters. Prior to clustering, a form of 'normalization' was carried out to convert PETs from one strand to the other, so that all discordant PETS were perceived to have come from a single strand. For each discordant mapping, both the original mapping and its reverse complement mapping were considered. One of the two mappings was selected according to following hierarchy of preferences. i) Mapping with lower chromosome value for 5' was chosen. ii) If both chromosomes were same, mapping which results in 5' tag mapping to '+' strand was chosen. iii) If no such mapping existed, mapping which resulted in 5' tag having smaller coordinate was chosen.

To cluster different dPETs which span the same fusion point, the following procedure was applied: the mapping location of the 5' and 3' tags of a given dPET was extended by the maximum insert size of the respective genomic library in both directions to create 5' and 3' searching windows. If the 5' and 3' tags of a second dPET mapped within the 5' and 3' searching window of the first dPET, the two PETs were defined as a cluster of the count 2 and the 5' and 3' searching windows were adjusted so that they contained the tag extensions (by the maximum library size) of the second dPET. dPETs which subsequently mapped with their 5' and 3' tags within the 5' and 3' searching windows, respectively, were assigned to this cluster and the windows were adjusted, if necessary. The number of dPETs clustering

together around a fusion point was represented by the cluster count. The genomic region which was covered by the 5' tags of a cluster was defined as the 5' anchor and the genomic region which was covered by the 3' tags of a cluster was defined as the 3' anchor. ***Note: PET sequencing analysis was performed by Pramila Ariyaratne, Genome Institute of Singapore, Singapore.***

Identification of structural variations

The characteristics of the dPET clusters were used to determine individual classes of structural variations. dPET clusters with an insert size which was larger than the size of the genomic library were defined as deletions. Tandem duplications were characterized by clusters where the 3' tags preceded the 5' tags of a mapped cluster (wrong ordering). Inversions were identified by two clusters of the category 'same chromosome, different strand' where the genomic regions which were covered by the clusters overlapped. Different kinds of insertions were defined by two dPET clusters which indicated that a genomic fragment (donor) has been inserted in a certain position (recipient). The two anchor regions on the recipient DNA strand were allowed to be at most twice the maximum library insert size apart from each other. Isolated translocations were characterized by clusters where 5' and 3' tags of a cluster mapped on different chromosome. Balanced translocations were identified by two isolated translocations clusters and the two genomic regions exchanged.

Superclustering

To determine the neighborhood of a breakpoint, the start and end points of each dPET cluster anchor region were extended by the maximum insert size of the respective genomic library as search windows. If windows of neighboring clusters overlapped with each other, dPET

clusters were grouped together into a supercluster. The procedure allowed an indirect connection of cluster A via B to C. The number of dPET clusters that could be joined together into a supercluster was represented by the supercluster size. In cases where >3 dPET clusters were interconnected, the breakpoint pairs were classified as 'complex' (intra- and inter-chromosomal). This method had two benefits: 1) it allowed the separation of complex regions and allowed a more reliable SV calling for the non-complex SVs, and 2) it gave an overview of which rearrangement points (and thereby regions) collectively formed an amplification. Since an amplified region can contain a number of different loci, its joint nature is neither obvious by looking at the copy number itself nor by investigating individual dPET clusters.

Comparison of libraries with different insert sizes

The comparison of dPET clusters across different insert size libraries was performed based on an overlap of the 5' and 3' anchor region extended by the individual library insert size. We started from the 10 kb library in MCF-7 and HCT116 and the 20 kb library in K562. For any given 10 kb or 20 kb isolated dPET cluster (supercluster count ≤3), the 5' and 3' anchor regions of the cluster was extended by the maximum length of the library towards the breakpoints to create a search window. If the 5' and 3' anchor regions of a dPET cluster from other insert size libraries which belonged to the same SV type fell into the search window, the clusters would be grouped as a common SV. If no other cluster could be found in the search window, the cluster would be categorized as a SV specific to that insert size library. ***Note: SVs identification, superclustering and different insert size libraries cross-comparison were performed by Charlie Lee, Genome Institute of Singapore, Singapore.***

Breakpoint confirmation by genomic PCR and Sanger sequencing

We validated a subset of breakpoints using genomic PCR. Primers were designed to span the breakpoint predicted by dPET clusters using repeat–masked human genome assembly (March 2006 assembly, Build 36). The maximum PCR product was 10 kb. PCR was carried out with JumpStart^{TM} REDAccuTaq LA DNA Ploymerase (Sigma-Aldrich) in a 50 µl reaction volume and with 20 ng of genomic DNA as the template. The following program was used: 1) Initial denaturation at 96°C for 30 sec, 2)15 cycles of 15 sec at 94°C, 30 sec at 58°C, 10 min at 68°C, 3) 25 cycles of 15 sec at 94°C, 30 sec at 55°C, 10 min at 68°C, 4) 68°C for 20 min. Fragments up to 10 kb in size were visualized by agarose gel electrophoresis. PCR products with single band at the expected size range were purified by QIAquick PCR Purification Kit (Qiagen), sequenced by conventional Sanger capillary methods and the resulting sequences were aligned to the reference sequence to identify breakpoints. We then compared breakpoint coordinates from Sanger sequencing with the breakpoint coordinates predicted by dPET clusters and determined a median resolution for each library size.

Copy number analysis

For estimation of copy number, the genome was divided into non-overlapping windows of equal size, which could be set as a parameter and depended on the overall coverage of the sequenced data. To determine the different mappability of reads in each window due to varying uniqueness across the genome, paired-end-tags of 1x genome coverage were simulated by creating randomly distributed fragments along the reference genome with size distribution similar to the actual experimental library and extracting the tags at both ends of the fragments. Sequencing and ligation errors of 1% and 5%, respectively, were also added to generate the simulated library. The simulated tags were then mapped to the reference genome

and paired using the same Applied Biosystems SOLiD mapping and pairing pipeline. Subsequently, the tag density in each window was corrected for its mappability by computing the ratio of number of mapped tags of the experimental library to the simulated library within the window. cPET tags were used for copy number estimation. Windows overlapping clusters of dPET tags were omitted for copy number estimation because the cPET tag count in these windows was no longer representative of the copy number.

The tag density in each window was corrected for GC bias according to the GC bias distribution of the DNA fragments observed in the DNA-PET library. The bias was calculated by taking the ratio of proportion of cPETs of each experimental library to the simulated library of corresponding size for each GC range of every 1 % (Figure 5.1). The variation in GC bias across the libraries is likely to be related to slight condition variations in the library construction procedure. Thus, for each DNA-PET library, the tag density in each 10 kb window of a given GC content was corrected for its bias based on the respective GC content bin. To estimate the copy number independently of the availability of a matched sample, we normalized the corrected tag density such that the median copy number was two copies. This is a reasonably valid assumption, as we would expect most parts of the genome to be normal. Due to significant amount of noise as a result of non-uniform sampling throughout the genome, the copy number estimates in the windows were smoothened to identify copy number segments and change points using a binary circular segmentation algorithm (SD=2) originally developed for aCGH data (Venkatraman et al. 2007). ***Note: copy number analysis was performed by Woo Xing Yi, Genome Institute of Singapore, Singapore.***

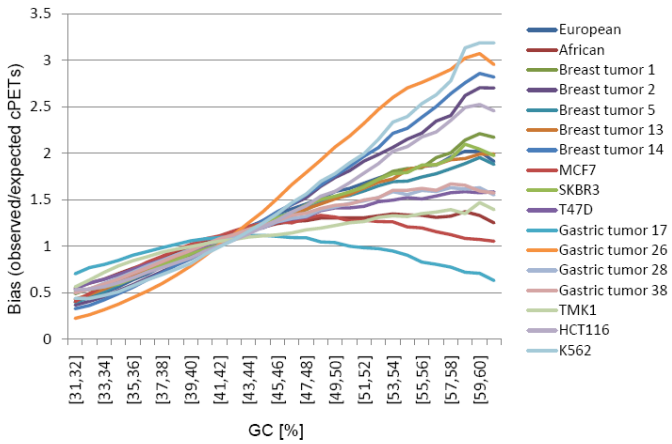

Figure 5.1. Distribution of GC bias of 17 DNA-PET samples.

The ratio of the observed numbers of cPETs to the simulated library (y-axis) within 10 Kb windows is used to correct the predicted copy number according to the GC content of each 10 Kb window (x-axis).

Fluorescence in situ hybridization (FISH)

FISH was performed as described by Koichiro (Inaki et al. 2011). In brief, nuclei were harvested by treating cells with 0.75M KCl for 20 min at 37°C and dropped on slides after few fixations. BAC DNA probes were labeled by nick translation in the presence of biotin-16-dUTP or digoxigenin-dUTP using Nick Translation System (Invitrogen). Prior to hybridization, slides were treated with 0.01% pepsin at 37°C for 5 min followed by 1 X PBS rinse, 1% formaldehyde 10 min treatment, 1 X PBS rinse (5 min) and dehydration through ethanol series (70%, 80%, and 100%). Denaturated probes were applied to these pre-treated slides and co-denaturated at 75°C for 5 min and hybridized at 37°C overnight. After post-hybridization washes and blocking, slides were revealed with avidin-conjugated fluorescein isothiocyanate (FITC) (Vector Laboratories). Slides were mounted with vectashield and

observed under epifluorescence microscope. ***Note: FISH experiment was performed by Valere group, Genome Institute of Singapore, Singapore.***

Reconstruction of genome structure by fusion point guided concatenation

Segmenting of the reference genome into contigs was done on the basis of breakpoints identified by dPET clusters and by identifying additional breakpoints with no physical cPET coverage. Contigs consecutive on the reference genome were then connected by a reference edge in the presence of connecting cPETs. Correspondingly, contigs linked by dPET clusters were represented by dPET edges where the edges were weighted by the count of the cluster. Locally amplified regions were then identified in the following way: firstly, the dPET edge with the highest weight was selected and the adjacent contigs to this edge were added to the amplicon graph. Then, for each contig in the graph, its neighbors were also added using both reference and dPET links as long as the neighbors were considered amplified (cPET estimated copy-number greater than 2). An amplicon graph was grown until no more contigs could be added in this fashion. The process was then repeated on the unused dPET edges, till none remained, resulting in a set of local amplicon graphs and only graphs with more than two contigs were considered further. ***Note: fusion point guided concatenation analysis was performed by Gao Song and Niranjan Nagarajan, Genome Institute of Singapore, Singapore.***

Materials and Methods for Chapter 3

Note: The Materials and Methods for Chapter 2 and 3 have a number of overlaps; where is occurs, Chapter 3 refers to Chapter 2, and the description in Chapter 2 includes slight modifications used in Chapter 3.

Cell culture

The human cell lines, SKBR3 and T47D were obtained from the American Type Culture Collection (ATCC), TMK1 was kindly provided by Dr Y. Ito (National University of Singapore; Agency for Science, Technology and Research, Singapore).

Clinical tumor samples

Tissue samples were obtained from five patients who had undergone surgery for breast cancer at the University Hospital Stockholm, Sweden, and from four patients who had undergone surgery for gastric cancer at the National University Hospital of Singapore. All breast tumors were from European patients and belonged to the basal-like subgroup based on micro array expression data. These samples were anonymized prior to sequencing and analysis therefore no clinical data are available. The gastric cancer specimens were from four male patients with advanced-stage gastric cancer (TNM stage 3a –gastric tumor 28, stage 3b –gastric tumor 26, stage 4 –gastric tumors 17 and 38, respectively) of Chinese (gastric tumors 17, 28, 38) and Malay (gastric tumor 26) ethnicity. Histologically, gastric tumor 26 was a Lauren classification diffuse-type, poorly differentiated signet ring cell carcinoma, while gastric tumors 17 and 38 were Lauren classification mixed-type, poorly-differentiated signet ring cell carcinomas, and gastric tumor 28 was a Lauren classification intestinal-type moderately-differentiated adenocarcinoma.

Genomic DNA extraction

The genomic DNA of cell lines was extracted by Blood & Cell Culture DNA Kits (Qiagen) and DNA of tumor samples was extracted using AllPrep DNA/RNA Mini Kit (Qiagen) according to the manufacturer's instruction.

DNA-PET library construction, sequencing and mapping

All libraries except gastric tumor 17 were constructed by the method described in chapter 2. The gastric tumor 17 was constructed by Long-Mate-Paired Library Construction method. Similar as the *EcoP15I* method, but LMP CAP adaptors (Applied Biosystems) with only a single 5' phosphorylated end were ligated to the hydro-sheared DNA, thus creating a nick on each strand after circularization of the DNA. Both nicks were translated >50 bp into the circularized genomic DNA fragment by DNA polymerase I, and paired-end tags of >50 bp were released by T7 exonuclease and S1 nuclease. SOLiD sequencing adaptors (Applied Biosystems) were ligated to the fragments and the ligated mixture were amplified with SOLiD PCR primers (Applied Biosystems) by a 13 cycles PCR. Finally 250-300 bp fragments were selected to generate mate paired sequencing libraries with average target genomic DNA on each end around 90bp by purified from PAGE gel and use as sequence template. The 2 x 50 bp sequencing was performed exactly according to the SOLiD 3 system Instrument Operation Guide and using the reagents from Applied Biosystems.

For all samples, one DNA-PET library was constructed and sequenced on one slide of SOLiD v2, except for Breast tumors 5, 13, and 14 for which two slides were sequenced. Gastric tumor 17 was sequenced on one SOLiD v3 slide. For MCF-7 and K562, two libraries were constructed and for HCT116, three libraries were constructed and the sequencing data was combined for downstream analysis. For K562, the second library was constructed with

Solexa adapters and was sequenced by nine lanes 2 x 36 bp of Solexa (Genome Analyzer II). The Solexa data was trimmed to 27 bp. All sequences have been submitted to the Gene Expression Omnibus (GEO) data base at National Center for Biotechnology Information (NCBI).

Sequence tags were mapped to the human reference sequence (NCBI Build 36), allowing two color-code mismatches for 25-bp reads and six mismatches for 50-bp reads and paired using the SOLiD System Analysis Pipeline Tool, Corona Lite (Applied Biosystems). ***Note: libraries IHB005, IHB013, IHB014, IHS012, IHT008, IHH027, DHG003, IHG005, IHG006, IHG009 and IHT009 were constructed by Audrey Teo. Genome Institute of Singapore, Singapore.***

Define dPET cluster count

To define the dPET cluster count which provides enough confidence to call an SV, we used two approaches for two different categories of noise.

1) *False deletion calls based on an increase of stretched (long insert) PETs*

To account for the possibility of stretched PETs leading to deletion artifacts, we estimated the expected number of false-positives as follows: for each library we computed a "stretch rate", *S*, as the proportion of library PETs that is considered stretched. For a cluster count *c,* the probability of obtaining a cluster of that count from stretched PETs is then given by *Sc* and this can be Bonferroni-corrected by the number of dPET clusters reported for a library to get an estimate for the expected number of false-positive clusters. Based on this model, the expected number of false-positive dPET clusters of count 3 or higher was $<10^{-4}$ for all libraries (Table 5.1).

2) *Noise model for chimeric ligation and p-value calculation*

The noise in DNA-PET data is mainly from random ligation. The null hypothesis assumes that, each DNA fragment has an equal chance to ligate with any other fragment to form a chimeric DNA-PET in a random and independent manner. Under this random model, the number of DNA-PETs that link two DNA fragments follows a hyper-geometric distribution. The formula is provided in Equation 1. One advantage of this model is that both PET frequency and the enrichment of the anchors are taken into account.

$$\Pr(I_{A,B} \mid N, c_A, c_B) = \frac{\binom{c_A}{I_{A,B}}\binom{2N - c_A}{c_B - I_{A,B}}}{\binom{2N}{c_B}} \qquad (1)$$

Equation 1 considers a library with N DNA-PETs (where each DNA-PET corresponds to a pair of reads). Consider two DNA fragments R_A and R_B and let c_A and c_B to be the number of reads mapping on R_A and R_B, respectively, where $c_{A,}$ $c_B << N$. Equation 1 gives the probability of choosing $I_{A,B}$ ends from c_A ends of Region R_A to form $I_{A,B}$ DNA-PET between Region R_A and Region R_B , when c_B ends are randomly chosen from $2N$ ends as anchors in Region R_B. By this, we were able to compute a p-value to test if $I_{A,B}$, the number of DNA-PETs between R_A and R_B, were over-represented. The p-values were corrected by the Benjamini and Hochberg method for multiple hypothesis testing (Benjamini and Hochberg 1995) as false discovery rate (or Q-value).

dPET clusters with smaller PET counts are more likely to be generated by random, than clusters with the larger PET counts. As a result, we assumed that the clusters with smaller PET counts were noise generated from random ligations, while the dPET clusters with larger PET counts were true signals generated from real SVs. The Q-value was used to classify the noise and true signals. To determine a reasonable cutoff between noise and true dPET clusters, we estimated the Area-Under-Curve (AUC) from Receiver Operating Characteristic

(ROC) curve for dPET cluster size cutoffs 2 to 10 (Figure 5.1). PET cutoffs 2 and 3 have the largest AUC values. Based on these rationales, we used a general dPET cluster size ≥3 to identify SVs acknowledging higher false discovery rates (FDRs) for the smaller clusters. Individual FDRs for each cluster based on the hypergeometric model are given in Appendix Table). ***Note: this analysis was performed by Li Guoliang, Genome Institute of Singapore, Singpaore.***

Table 5.1: Expected numbers of long insert dPET clusters

Library name	Sample	Stretch rate[1)]	cPET span	Long span dPET cutoff[2)]	Number of dPET clusters	P-value[3)]	E-value[4)]
IHH027	European	0.00178019	7400-10977	12074	666	5.64E-09	3.76E-06
IHH022	African	0.00626796	7132-9854	10839	356	2.46E-07	8.77E-05
IHB001	Breast tumor 1	0.00168624	8323-12436	13679	313	4.79E-09	1.50E-06
IHB002	Breast tumor 2	0.00431883	7742-12163	13379	426	8.06E-08	3.43E-05
IHB005	Breast tumor 5	0.0064154	8109-12120	13332	242	2.64E-07	6.39E-05
IHB013	Breast tumor 13	0.0017668	4012-6325	6957	957	5.52E-09	5.28E-06
IHB014	Breast tumor 14	0.00474705	7592-12520	13772	434	1.07E-07	4.64E-05
IHM005006	MCF7	0.001928085	8099-16217	17838	1047	7.17E-09	7.50E-06
IHS012	SKBR3	0.00134481	7229-9639	10602	1145	2.43E-09	2.78E-06
IHT008	T47D	0.00571014	7816-11617	12778	376	1.86E-07	7.00E-05
DHG003	Gastric tumor 17	0.00184182	8630-11310	12441	1126	6.25E-09	7.04E-06
IHG009	Gastric tumor 26	0.00316919	8108-11653	12818	586	3.18E-08	1.87E-05
IHG005	Gastric tumor 28	0.00375142	8152-11786	12964	1238	5.28E-08	6.54E-05
IHG006	Gastric tumor 38	0.0029249	7561-11503	12653	550	2.50E-08	1.38E-05
IHT009	TMK1	0.00155585	8760-11920	13112	1253	3.77E-09	4.72E-06
IHH003020026	HCT116	0.00456048	7200-11780	12958	883	9.48E-08	8.38E-05
IHK006007	K562	0.00140964	6846-10248	11272	939	2.80E-09	2.63E-06

[1)] Proportion of PETs which are longer than the long span dPET cutoff
[2)] To define a long span dPET, the maximum cPET value is multiplied by 1.1
[3)] Probability of Sc, with S=stretch rate and c=cluster size (set to 3 for all libraries)
[4)] Expected number of false positive stretched clusters among the dPET clusters

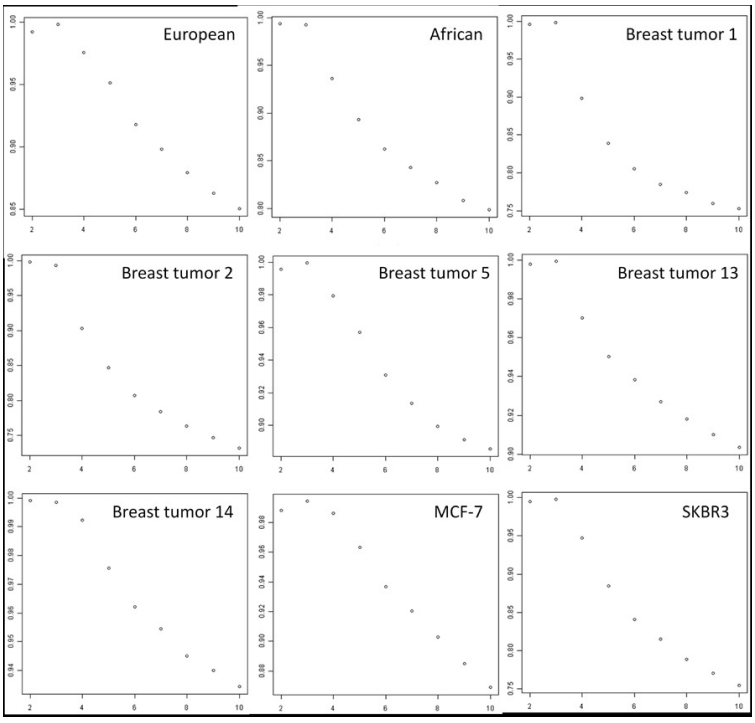

Figure 5.2. A. Area-Under-Curve (AUC) from Receiver Operating Characteristic (ROC) curve for dPET cutoffs 2 to 10 of nine DNA-PET samples.

dPET cluster size cutoffs are shown on the x-axis, AUC is shown on the y-axis. Highest AUC value indicates best model (here dPET cluster count cutoff) for highest true positive vs. false positive ratio.

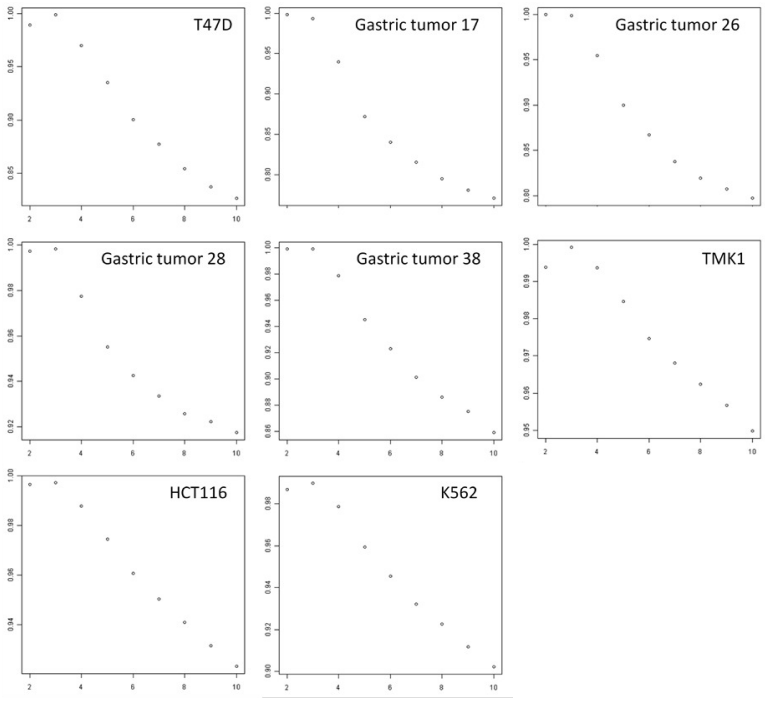

Figure 5.2. B. Area-Under-Curve (AUC) from Receiver Operating Characteristic (ROC) curve for dPET cutoffs 2 to 10 of eight DNA-PET samples.

dPET cluster size cutoffs are shown on the x-axis, AUC is shown on the y-axis. Highest AUC value indicates best model (here dPET cluster count cutoff) for highest true positive vs. false positive ratio.

SNP and sequencing error simulation for mapping and clustering

To investigate if and to what extent sequencing errors, SNPs, and point mutations result in false positive SV calls in our pipeline, we simulated 25 bp paired-end-tags of 3x (sequence) genome coverage by creating randomly distributed 9.5 kb fragments (with a typical size distribution) along the reference genome with a point mutation rate of 0.001 and sequencing errors taken from two sequenced libraries (IHH027 [European] and IHT009 [TMK1]). The two simulated libraries gave 120-fold and 123-fold physical coverage, respectively, which was higher than the average coverage of 81-fold for the 17

analyzed genomes. The two simulations resulted in 128 and 133 dPET clusters, respectively, which passed our quality criteria. Since the SV calling is dependent on the coverage, the absolute numbers of artificial clusters is expected to vary according to the physical coverage. In the 17 genomes, we identified on average 737 dPET clusters (666 and 356 in the two normal control samples). This suggested that ca. 18% (133/737) of the 15 predicted SVs in this study are false positives due to SNPs, point mutations, and sequencing errors. We intersected the dPET clusters of the two simulations with the SV predicting clusters of the 17 analyzed genomes. Of the simulation based clusters, 27 and 29, respectively, matched DNA-PET clusters of the 17 genomes. We indicated these clusters in Appendix Tables as artifacts ('sim10kb25bpIHH027' and 'sim10kb25bpIHT009'). ***Note: this analysis was performed by Niranjan Nagarajan, Genome Institute of Singapore, Singpaore.***

Cross-genome comparison

Comparison of clusters across different genomes was performed based on an overlap of the 5' and 3' anchor regions extended by 10 kb on both sides. If the 5' anchor region of a cluster from a second library was overlapping with the 5' extended anchor region of a cluster of the first library and the same was true for the 3' anchor regions, the two clusters were grouped together and the 10 kb extension of the anchor regions were adjusted according to the outermost start and end anchor coordinates. Breakpoint locations of the pooled coordinates were used to compare the identified SVs with SVs in the database of genomic variants (http://projects.tcag.ca/variation/) (Iafrate et al. 2004), paired-end

sequencing studies of noncancer individuals (Kidd et al. 2008; Korbel et al. 2007) and paired-end sequencing data of 24 breast cancer genomes (Stephens et al. 2009). The fraction of an SV that overlapped with another event was calculated by the percentage of overlap relative to the larger event. Gene annotations was based on RegSeq Genes downloaded from UCSC (http://genome.ucsc.edu/) (Rhead et al. 2010) on May 14, 2009 using library-specific breakpoints. ***Note: this analysis was performed by Charlie Lee, Genome Institute of Singapore, Singapore.***

Sequence features at the breakpoints of SVs

To investigate for the presence of potential sequence homology between paired breakpoints, we extracted sequences 10 kb up- and downstream around predicted 20 breakpoints from the human reference genome at University of California-Santa Cruz (UCSC) Genome Bioinformatics (hg18, http://genome.ucsc.edu/, (Kent et al. 2002). The pairs of 20 kb sequences associated with each dPET cluster were then aligned by BLAST (bl2seq, (Altschul et al. 1990). This usually resulted in multiple alignments and we used the one with the highest score (defined as "BLAST score"). The distribution of BLAST scores obtained was compared with the distribution obtained from 1,000 randomly chosen 20 kb segments of hg18. Less than 3.25% of the random pair alignments resulted in BLAST scores >300 (data not shown) and this threshold was selected to identify pairs of breakpoints with “significant” homology. These regions with significant homology were further annotated using genomic features from the UCSC genome browser such as segmental duplications (SDs) and L1 repeats. ***Note: this analysis was performed by Zhao Hao, Genome Institute of Singapore, Singapore.***

Sequence similarity of breakpoint pairs by BLAST

A dilemma of genome studies using short read sequencing data is that two genomic regions A and B, which show sequence homology, can cause mapping artifacts if reads of region A map better to region B due to errors in sequencing or in the reference sequence or due to the presence of SNPs. On the other hand, it seems unwise to ignore all dPETs which connect two regions with high sequence homology since sequence homology is considered a possible mechanism for NAHR (Gu et al. 2008). We therefore annotated the dPET clusters with an additional pair-wise Blast analysis (BlastScore1 in Appendix Table) in which the 5' anchor region of a dPET cluster plus a 15 kb extension toward the breakpoint was aligned to the paired 3' anchor region plus 15 kb extension (this is a modification of the Blast analysis shown in Figure. 3.19 where 10 kb up- and downstream of each breakpoint have been aligned). This annotation may be used to prioritize SVs for downstream validation of predicted fusion genes (Inaki et al. 2011). SVs with high sequence homology are difficult to validate by PCR due to hindered unique primer design and possible amplification artifacts based on cross-priming of heterogeneous extension products. Twenty-one percent of all SVs (2,635/12,537) had a sequence homology between their breakpoint regions which resulted in a Blast score ≥1,000, whereas 15% of the cancer SVs (957/6,410) showed high sequence homology.

Note: this analysis was performed by Zhao Hao, Genome Institute of Singapore, Singapore.

Copy number of dPET clusters

It is desirable to connect a copy number value with a dPET cluster count. This, however, cannot be achieved by the same method which we used for copy number estimation based on the cPET tags, since the mapping of dPET is more stringent in the Corona Lite SOLiD pipeline (a process termed 'rescuing' favors paired mapping of cPETs). Despite these limitations, it is possible to use the cPET based copy number information within an amplified region and correlate it with the dPET cluster counts of this region. The copy number of the high copy regions of the *BMP7* and *ZNF217* loci in MCF-7 can be estimated at 47 and 32 copies, respectively, based on the cPET tags. Both regions are connected by the largest dPET cluster of size 1,176 (TD 1,176). Considering the region with the lower copy number and subtracting arbitrarily 2 copies of the hypertriploid to hypotetraploid cell line as not connected by this cluster, there might be 30 copies of the rearrangement point correlated with the cluster size of 1,176 (ca. 39 dPETs per copy for this particular region). This suggests that ca. 25 of the 30 copies of TD 1,176 (Figure 3.23) were pasted in the four genomic surroundings illustrated in Figure 3.23. ***Note: this analysis was performed by Woo Xing Yi, Genome Institute of Singapore, Singapore.***

Quantitative polymerase chain reaction (qPCR)

qPCR was performed as described by Inaki et al. (Inaki et al. 2011). Primers used were as follows: 5'-TCCCTGGAGAAGAGCTACGA-3'and 5'-AGGAAGGAAGGCTGGAAGAG-3' for ACTB (*ACTB*), 5'-

GCCTGCAAGATAGCCATTTC-3' and 5'-TGGGTGGAAGAATTCCTTGT-3' for *BMP7*, 5'-TCTGACCCAACAGTCCC-3' and 5'-CTGGAGACAAGGGATTTCC-3' for *ZNF217* (long), 5'-GGAAGGTGGTTCTGAAGACG-3' and 5'-TGAACGGAAAAACTTTCCACA-3' for *ZNF217* (short). Each Ct value was subtracted by *ACTB* Ct value (dCt) and relative expression level was calculated as $2^{(-dCt)}$. ***Note: this analysis was performed by Koichrio Inaki, Genome Institute of Singapore, Singapore.***

Statistical analysis

A two-tailed x^2 test was used to test for differences between the fractions of cancer and normal breakpoints with sequence homologies, and to test whether the proportion of normal and cancer breakpoints in gene deserts, genes and regulatory regions was in accordance with the size proportion of the respective regions relative to the human genome. Mann-Whitney U Test was applied to compare SV size distributions between normal samples and the different cancer categories and to test for differences between the frequency of short unpaired inversions per chromosome of breast and gastric cancer genomes. ***Note: this analysis was performed by Charlie Lee, Genome Institute of Singapore, Singapore.***

References

Adeyinka, A. et al. 2000. Spectral karyotyping and chromosome banding studies of primary breast carcinomas and their lymph node metastases. *Int J Mol Med* **5:** 235-240.

Altschul, S.F. et al. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

Aparicio, S.A. et al. 2010. Does massively parallel DNA resequencing signify the end of histopathology as we know it? *J Pathol* **220:** 307-315.

Armour, J.A. et al. 2000. Measurement of locus copy number by hybridisation with amplifiable probes. *Nucleic Acids Res* **28:** 605-609.

Ashley, E.A. et al. 2010. Clinical assessment incorporating a personal genome. *Lancet* **375:** 1525-1535.

Bajpai, S. et al. 2009. Loss of alpha-catenin decreases the strength of single E-cadherin bonds between human cancer cells. *J Biol Chem* **284:** 18252-18259.

Banerji, S. et al. 2012. Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486:** 405-409.

Bashir, A. et al. 2008. Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. *PLoS Comput Biol* **4:** e1000051.

Bentley, D.R. et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456:** 53-59.

Beroukhim, R. et al. 2010. The landscape of somatic copy-number alteration across human cancers. *Nature* **463:** 899-905.

Bignell, G.R. et al. 2010. Signatures of mutation and selection in the cancer genome. *Nature* **463:** 893-898.

Cahill, D. et al. 2006. Mechanisms of eukaryotic DNA double strand break repair. *Front Biosci* **11:** 1958-1976.

Campbell, P.J. et al. 2008. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* **40:** 722-729.

Carey, L.A. et al. 2006. Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *JAMA* **295:** 2492-2502.

Chaisson, M.J. et al. 2009. De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome Res* **19:** 336-346.

Chen, J. et al. 2008. Scanning the human genome at kilobase resolution. *Genome Res* **18:** 751-762.

Clark, M.J. et al. 2010. U87MG decoded: the genomic sequence of a cytogenetically aberrant human cancer cell line. *PLoS Genet* **6:** e1000832.

Cordaux, R. et al. 2009. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10:** 691-703.

Curtis, C. et al. 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486:** 346-352.

Daley, G.Q. et al. 1990. Induction of chronic myelogenous leukemia in mice by the P210bcr/abl gene of the Philadelphia chromosome. *Science* **247:** 824-830.

Ding, L. et al. 2010. Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464:** 999-1005.

Drmanac, R. et al. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327:** 78-81.

Edwards, P.A. 2010. Fusion genes and chromosome translocations in the common epithelial cancers. *J Pathol* **220:** 244-254.

Ellis, M.J. et al. 2012. Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* **486:** 353-360.

Feuk, L. et al. 2006. Structural variation in the human genome. *Nat Rev Genet* **7:** 85-97.

Friedman, J.M. 2009. High-resolution array genomic hybridization in prenatal diagnosis. *Prenat Diagn* **29:** 20-28.

Fujimoto, A. et al. 2010. Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. *Nat Genet* **42:** 931-936.

Fullwood, M.J. et al. 2009. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462:** 58-64.

Garcia M, J.A., Ward EM, Center MM, Hao Y, Siegel RL, Thun MJ. 2007. Global Cancer Facts and Figures 2007. Atlanta, GA: American Cancer Society.

Groffen, J. et al. 1984. Philadelphia chromosomal breakpoints are clustered within a limited region, bcr, on chromosome 22. *Cell* **36:** 93-99.

Gu, W. et al. 2008. Mechanisms for human genomic rearrangements. *Pathogenetics* **1:** 4.
Guffanti, A. et al. 2009. A transcriptional sketch of a primary human breast cancer by

454 deep sequencing. *BMC Genomics* **10:** 163.

Hampton, O.A. et al. 2009. A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome. *Genome Res* **19:** 167-177.

Hanahan, D. et al. 2000. The hallmarks of cancer. *Cell* **100:** 57-70.

Heisterkamp, N. et al. 1990. Acute leukaemia in bcr/abl transgenic mice. *Nature* **344:** 251-253.

Hicks, J. et al. 2006. Novel patterns of genome rearrangement and their association with survival in breast cancer. *Genome Res* **16:** 1465-1479.

Hillmer, A.M. et al. 2011. Comprehensive long-span paired-end-tag mapping reveals characteristic patterns of structural variations in epithelial cancer genomes. *Genome Res* **21:** 665-675.

Holt, R.A. et al. 2008. The new paradigm of flow cell sequencing. *Genome Res* **18:** 839-846.

Hughes, T. et al. 2006. Monitoring CML patients responding to treatment with tyrosine kinase inhibitors: review and recommendations for harmonizing current methodology for detecting BCR-ABL transcripts and kinase domain mutations and for expressing results. *Blood* **108:** 28-37.

Iafrate, A.J. et al. 2004. Detection of large-scale variation in the human genome. *Nat Genet* **36:** 949-951.

Inaki, K. et al. 2011. Transcriptional consequences of genomic structural aberrations in breast cancer. *Genome Res* **21:** 676-687.

Jonsson, G. et al. 2007. High-resolution genomic profiles of breast cancer cell lines assessed by tiling BAC array comparative genomic hybridization. *Genes Chromosomes Cancer* **46:** 543-558.

Kent, W.J. et al. 2002. The human genome browser at UCSC. *Genome Res* **12:** 996-1006.

Kidd, J.M. et al. 2008. Haplotype sorting using human fosmid clone end-sequence pairs. *Genome Res* **18:** 2016-2023.

Korbel, J.O. et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318:** 420-426.

Kozlowski, P. et al. 2008. New applications and developments in the use of multiplex ligation-dependent probe amplification. *Electrophoresis* **29:** 4627-4636.

Krzywinski, M. et al. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res* **19:** 1639-1645.

Kuttler, F. et al. 2007. Formation of non-random extrachromosomal elements during development, differentiation and oncogenesis. *Semin Cancer Biol* **17:** 56-64.

Leary, R.J. et al. 2010. Development of personalized tumor biomarkers using massively parallel sequencing. *Sci Transl Med* **2:** 20ra14.

Lee, W. et al. 2010. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465:** 473-477.

Ley, T.J. et al. 2008. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456:** 66-72.

Lin, C. et al. 2009. Nuclear receptor-induced chromosomal proximity and DNA breaks underlie specific translocations in cancer. *Cell* **139:** 1069-1083.

Link, D.C. et al. 2011. Identification of a novel TP53 cancer susceptibility mutation through whole-genome sequencing of a patient with therapy-related AML. *JAMA* **305:** 1568-1576.

Liu, G. et al. 2009. Genomic amplification and oncogenic properties of the GASC1 histone demethylase gene in breast cancer. *Oncogene* **28:** 4491-4500.

Lo, A.W. et al. 2002. Chromosome instability as a result of double-strand breaks near telomeres in mouse embryonic stem cells. *Mol Cell Biol* **22:** 4836-4850.

Mardis, E.R. et al. 2009. Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med* **361:** 1058-1066.

Margulies, M. et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376-380.

McKernan, K.J. et al. 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res* **19:** 1527-1541.

Meyerson, M. et al. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet* **11:** 685-696.

Meyerson, M. et al. 2010. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet* **11:** 685-696.
Morin, R. et al. 2008. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* **45:** 81-94.

Murnane, J.P. et al. 2004. Chromosome rearrangements resulting from telomere dysfunction and their role in cancer. *Bioessays* **26:** 1164-1174.

Nagarajan, N. et al. 2009. Parametric complexity of sequence assembly: theory and applications to next generation sequencing. *J Comput Biol* **16:** 897-908.

Navin, N. et al. 2011a. Future medical applications of single-cell sequencing in cancer. *Genome Med* **3:** 31.

Navin, N. et al. 2011b. Tumour evolution inferred by single-cell sequencing. *Nature* **472:** 90-94.

Okuno, Y. et al. 2004. Structure of a palindromic amplicon junction implicates microhomology-mediated end joining as a mechanism of sister chromatid fusion during gene amplification. *Nucleic Acids Res* **32:** 749-756.

Pareek, C.S. et al. 2011a. Sequencing technologies and genome sequencing. *J Appl Genet* **52:** 413-435.

Pareek, C.S. et al. 2011b. Sequencing technologies and genome sequencing. *J Appl Genet*.

Pleasance, E.D. et al. 2010a. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463:** 191-196.

Pleasance, E.D. et al. 2010b. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463:** 184-190.

Plumb, C.L. et al. 2009. Modulation of the tumor suppressor protein alpha-catenin by ischemic microenvironment. *Am J Pathol* **175:** 1662-1674.

Raphael, B.J. et al. 2008. A sequence-based survey of the complex structural organization of tumor genomes. *Genome Biol* **9:** R59.

Rhead, B. et al. 2010. The UCSC Genome Browser database: update 2010. *Nucleic Acids Res* **38:** D613-619.

Rowley, J. 1973. A new constistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature* **243:** 290-293.

Ruan, Y. et al. 2007. Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). *Genome Res* **17:** 828-838.

Santarius, T. et al. 2010. A census of amplified and overexpressed human cancer genes. *Nat Rev Cancer* **10:** 59-64.

Schadt, E.E. et al. 2010. A window into third-generation sequencing. *Hum Mol Genet* **19:** R227-240.

Schouten, J.P. et al. 2002. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res* **30:** e57.

Sellner, L.N. et al. 2004. MLPA and MAPH: new techniques for detection of gene deletions. *Hum Mutat* **23:** 413-419.

Shah, S.P. et al. 2009. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461:** 809-813.

Shah, S.P. et al. 2012. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486:** 395-399.

Shaikh, T.H. et al. 2009. High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. *Genome Res* **19:** 1682-1690.

Sharp, A.J. et al. 2006. Structural variation of the human genome. *Annu Rev Genomics Hum Genet* **7:** 407-442.

Soda, M. et al. 2007. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature* **448:** 561-566.

Speicher, M.R. et al. 2005. The new cytogenetics: blurring the boundaries with molecular biology. *Nat Rev Genet* **6:** 782-792.

Stefansson, H. et al. 2005. A common inversion under selection in Europeans. *Nat Genet* **37:** 129-137.

Stephens, P.J. et al. 2009. Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462:** 1005-1010.

Stephens, P.J. et al. 2012. The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486:** 400-404.

Stratton, M.R. et al. 2009. The cancer genome. *Nature* **458:** 719-724.

Tanaka, H. et al. 2009. Palindromic gene amplification--an evolutionarily conserved role for DNA inverted repeats in the genome. *Nat Rev Cancer* **9:** 216-224.

Tomlins, S.A. et al. 2005. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310:** 644-648.

Tuzun, E. et al. 2005. Fine-scale structural variation of the human genome. *Nat Genet* **37:** 727-732.

Tyson, J. et al. 2009. Quadruplex MAPH: improvement of throughput in high-resolution copy number screening. *BMC Genomics* **10:** 453.

Venkatraman, E.S. et al. 2007. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23:** 657-663.

Volik, S. et al. 2006. Decoding the fine-scale structure of a breast cancer genome and transcriptome. *Genome Res* **16:** 394-404.

Volpi, E.V. et al. 2008. FISH glossary: an overview of the fluorescence in situ hybridization technique. *Biotechniques* **45:** 385-386, 388, 390 passim.

Wang, K. et al. 2011. Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. *Nat Genet* **43:** 1219-1223.

Wei, C.L. et al. 2006. A global map of p53 transcription-factor binding sites in the human genome. *Cell* **124:** 207-219.

Welch, J.S. et al. 2011. Use of whole-genome sequencing to diagnose a cryptic fusion oncogene. *JAMA* **305:** 1577-1584.

Wetzel, J. et al. 2011. Assessing the benefits of using mate-pairs to resolve repeats in de novo short-read prokaryotic assemblies. *BMC Bioinformatics* **12:** 95.

Wu X M, X.H.S. 2009. Progress in the detection of human genome structural variations. *Sci China Ser C-Life Sci* **52:** 560-567.

Xie, C. et al. 2009. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* **10:** 80.

Xu, J. et al. 2003. Advances in molecular cytogenetics for the evaluation of mental retardation. *Am J Med Genet C Semin Med Genet* **117C:** 15-24.

Zang, Z.J. et al. 2012. Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes. *Nat Genet* **44:** 570-574.

Zhang, J. et al. 2011. The impact of next-generation sequencing on genomics. *J Genet Genomics* **38:** 95-109.

Appendices

Appendices include all the raw data in excel tables and PDF files. All appendices, a PDF version of this thesis, and published paper related to this thesis are included in attached CD-ROM.

Thank you for reading!